Exhibit M of Rockett Declaration
with Response dated 04/16/04
In USSN: 09/828,423

JOURNAL OF
BIOLUMINESCENCE AND CHEMILUMINESCENCE

Bioluminescence and Chemiluminescence: Studies and Applications in Biology and Medicine

Proceedings of the Vth International
Symposium on Bioluminescence and
Chemiluminescence

Editors:
M. Pazzagli, E. Cadenas, L. J. Kricka,
A. Roda and P. E. Stanley

Volume 4 1989

WILEY
Chichester · New York · Brisbane · Toronto · Singapore

JBCHE7 4(1) 1-64(
ISSN 0884-3996

CLIN. CHEM. 37/11, 1955–1967 (1991)

Multianalyte Microspot Immunoassay—Microanalytical "Compact Disk" of the Future

R. P. Ekins and F. W. Chu

Throughout the 1970s, controversy centered both on immunoassay "sensitivity" per se and on the relative sensitivities of labeled antibody (Ab) and labeled analyte methods. Our theoretical studies revealed that RIA sensitivities could be surpassed only by the use of very high-specific-activity nonisotopic labels in "noncompetitive" designs, preferably with monoclonal antibodies. The time-resolved fluorescence methodology known as DELFIA—developed in collaboration with LKB/Wallac—represented the first commercial "ultrasensitive" nonisotopic technique based on these theoretical insights, the same concepts being subsequently adopted in comparable methodologies relying on the use of chemiluminescent and enzyme labels. However, high-specific-activity labels also permit the development of "multianalyte" immunoassay systems combining ultrasensitivity with the simultaneous measurement of tens, hundreds, or thousands of analytes in a small biological sample. This possibility relies on simple, albeit hitherto-unexploited, physicochemical concepts. The first is that *all* immunoassays rely on the measurement of Ab occupancy by analyte. The second is that, provided the Ab concentration used is "vanishingly small," fractional Ab occupancy is independent of both Ab concentration and sample volume. This leads to the notion of "ratiometric" immunoassay, involving measurement of the ratio of signals (e.g., fluorescent signals) emitted by two labeled Abs, the first (a "sensor" Ab) deposited as a microspot on a solid support, the second (a "developing" Ab) directed against either occupied or unoccupied binding sites of the sensor Ab. Our preliminary studies of this approach have relied on a dual-channel scanning-laser confocal microscope, permitting microspots of area 100 μm^2 or less to be analyzed, and implying that an array of 10^6 Ab-containing microspots, each directed against a different analyte, could, in principle, be accommodated on an area of 1 cm^2. Although measurement of such analyte numbers is unlikely ever to be required, the ability to analyze biological fluids for a wide spectrum of analytes is likely to transform immunodiagnostics in the next decade.

Additional Keyphrases: *ratiometric immunoassays · scanning-laser confocal microscope · fluoroimmunoassay*

Immunoassay and other protein-binding assay methods based on the use of radioisotopic labels have played a major role in medicine during the past three decades. Their utility and importance have derived primarily from the structural specificity of many reactions between binding proteins and analytes and the detectability of isotopically labeled reagents, the latter endowing such techniques with "exquisite sensitivity." Recently, however, interest has increasingly focused on nonisotopic techniques based on identical analytical principles, differing only in the nature of the marker used to label the reactant (e.g., antibody or antigen), whose distribution between reacted ("bound") and unreacted ("free") fractions constitutes the assay "response."

The basic aims underlying this interest can be broadly classed under four main headings:

- avoidance of the environmental, legal, economic, and practical disadvantages of isotopic techniques (e.g., limited shelf life of isotopically labeled reagents, problems of radioactive waste disposal, cost and complexity of radioisotope counting equipment), particularly those impeding the development of, for example, simple diagnostic kits for home or doctor's office use;
- achievement of greater assay sensitivity;
- "direct" measurement of analyte concentrations by use of transducer-based "immunosensors";
- simultaneous measurement of multiple analytes ("multianalyte assay").

In this presentation I will focus primarily on the last of these objectives, using this to set out the principles underlying our present attempts to develop a new "miniaturized" technology that will permit the simultaneous measurement of an unlimited number of analytes in a small biological sample such as a single drop of blood. However, retention (and, if possible, improvement) of the high sensitivities of conventional isotopic techniques is a basic aim not only of our own studies in this area but also of most other endeavors falling under the above headings. It is therefore appropriate to preface this paper with a discussion of the general principles underlying the attainment of high binding-assay sensitivity.

Immunoassay Sensitivity: Some Basic Concepts

Definition of Assay Sensitivity

The need to establish assay conditions yielding maximal sensitivity underlay the independent construction of mathematical theories of immunoassay design by both Yalow and Berson (*1*) and Ekins et al. (*2*) in the course of the original development of these methods in the early 1960s. Regrettably, these theoretical studies led to a prolonged controversy, arising largely from the conflicting concepts of "sensitivity" adopted by the two groups (see Figure 1). Briefly, Berson and Yalow, in their many publications relating to immunoassay design (e.g., *1*, *3*), defined sensitivity as the slope of the

Department of Molecular Endocrinology, University College and Middlesex School of Medicine, Mortimer St., London W1N 8AA, U.K.

Presented at the 23rd annual Oak Ridge Conference on Advanced Analytical Concepts for the Clinical Laboratory, St. Louis, MO, April 1991.

Received May 6, 1991; accepted August 20, 1991.

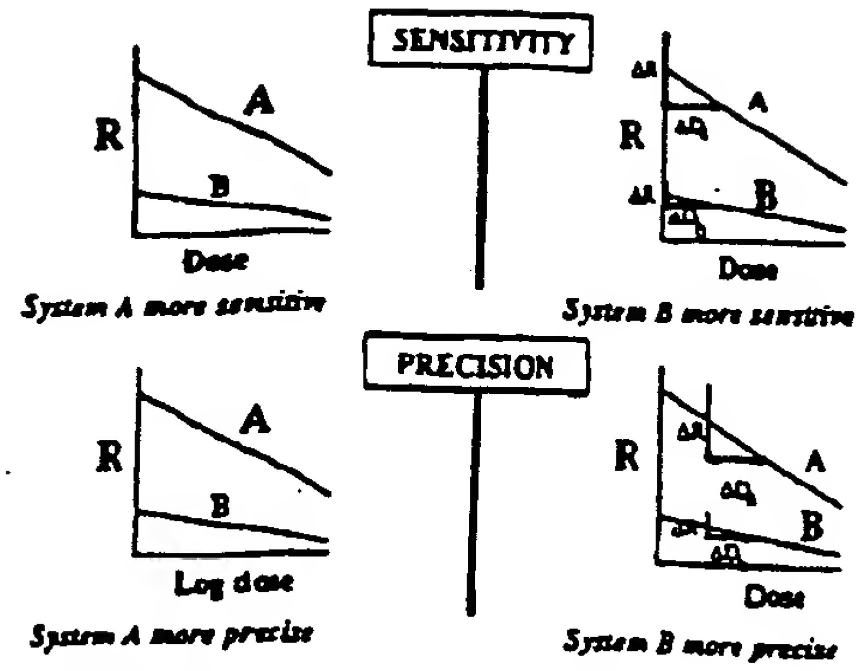


Fig. 1. The differing concepts of sensitivity and precision underlying radioimmunoassay design theories developed by (*left*) Yalow and Berson (e.g., *1, 3*) and (*right*) Ekins et al. (*2, 4*)

Yalow and Berson define assay A as more sensitive because it yields a response curve of greater slope. Ekins et al. define assay B as more sensitive because the imprecision of measurement of zero dose (σ_0) is less. Yalow and Berson likewise define an assay system as more precise if it yields a steeper response curve when data are plotted on a log dose scale

response curve relating the fraction or percentage of labeled antigen bound (b) to analyte concentration ([H]). In contrast, Ekins et al. (e.g., *2, 4*) defined sensitivity as the (im)precision of measurement of zero dose, this quantity being indicative of, and essentially equivalent to, the lower limit of detection.

The key difference between these two definitions clearly lies in the dependence of the assay detection limit on the error (imprecision) in the measurement of the response variable. By neglecting this crucial factor, the "response curve slope" definition leads to many obvious absurdities. For example, plotting conventional RIA data in terms of the response metameter B/F (i.e., the bound to free ratio) suggests that assay "sensitivity" is *increased* by increasing the antibody concentration in the system; however, the converse conclusion is reached if identical data are plotted in terms of F/B (see Figure 2). Observation of the shape and slopes of response curves without detailed error analysis thus constitutes a totally misleading guide to optimal immunoassay design. This approach has, however, characterized many of the studies conducted in the immunoassay field during the past 30 years, and has been the source of much mythology. For example, consideration of the Law of Mass Action reveals that, when response curves corresponding to different antibody concentrations are plotted in terms of b vs [H], the maximal slope at zero dose is obtained for a concentration of 0.5/*K* (where *K* is the affinity constant), in which circumstance the zero dose response (b_0) is 33%. This conclusion led to Berson and Yalow's enunciation of the well-known dictum (which, albeit erroneous, is broadly adhered to by many immunoassay practitioners and kit manufacturers) that, to maximize RIA sensitivity, the amount of antibody to use in the system is that which binds 33% of labeled antigen in the absence of unlabeled antigen (*1, 3*).

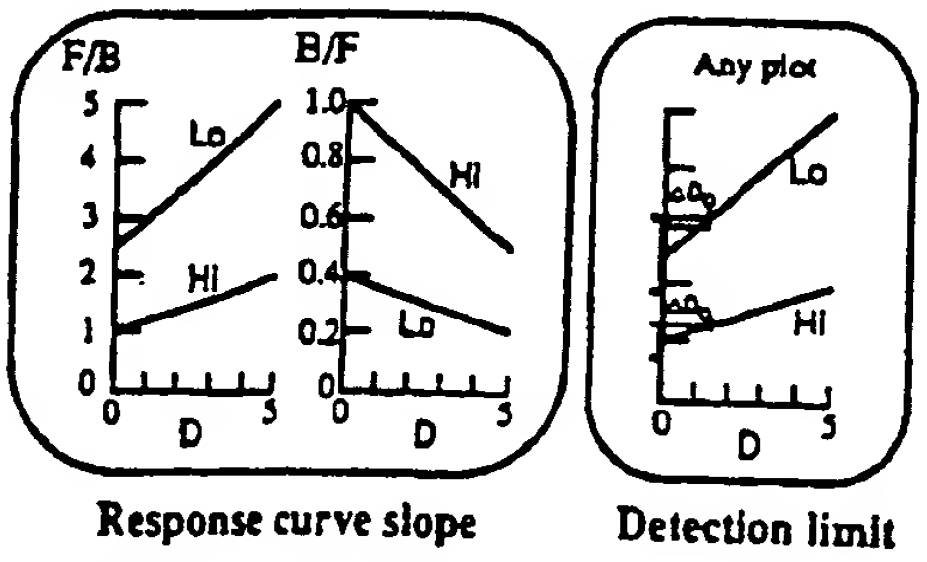


Fig. 2. Schematic representation of RIA dose–response curves observed for high and low antibody concentrations plotted in terms of (*left*) the free/bound fraction (F/B); (*center*) the bound/free fraction (B/F)

Note that the low antibody concentration yields a response curve of greater slope when the assay response is plotted in terms of F/B, but of lower slope when plotted in terms of B/F. The precision of measurement of zero dose (ΔD_0) is independent of the coordinate frame used to plot assay data (see *right*)

Disagreement regarding the concept of sensitivity inevitably led to prolonged dispute regarding immunoassay design (*5*). However, although it is still common to encounter publications in the field that rely solely on the response curve slope as a measure of sensitivity, the assay detection limit is now widely accepted as the only valid indicator of this parameter, and we do not therefore intend to dwell further on this issue here. It is nevertheless relevant to an understanding of the "miniaturized" assay methodology described below to emphasize that untenable concepts of both sensitivity *and* precision underlie many of the commonly accepted rules governing current immunoassay-design practice, some of which are contravened in our own approach.

Basic Immunoassay Designs

It is likewise important in the present context to comprehend the basis of the various types of immunoassays currently in use, and the constraints on the sensitivities of which they are potentially capable. The radioimmunoassay and analogous protein-binding assay techniques originally developed for the measurement of insulin by Yalow and Berson (*6*), and of thyroxin and vitamin B_{12} by Ekins and Barakat (*7, 8*), relied on the use of a labeled analyte marker to reveal the products of the binding reactions between analyte and binder (Figure 3, left). This approach has subsequently often been portrayed as relying on "competition" between labeled and unlabeled analyte molecules for a limited number of protein-binding sites, such assays being frequently referred to as "competitive."

Subsequently, Wide et al. in Sweden (*9*), followed shortly by Miles and Hales in the U.K. (*10*), developed labeled antibody methods (Figure 3, right). These methods represented an extension of the "labeled reagent" methods (utilizing radiolabeled organic compounds such as ^{131}I-labeled *p*-iodosulfonyl chloride, [^{3}H]acetic anhydride, and other similar reagents) devised, during the early 1950s, by Keston et al. (*11*), Avivi et al. (*12*), and others for quantifying amino acids, steroid and thyroid hormones, etc. Although radiolabeled antibody methods (immunoradiometric assays; IRMAs) were originally claimed (*13*) to be more sensitive than methods based on the use of radiolabeled analyte, these claims were supported by neither rigorous theoretical analysis nor persuasive experimental evidence, and for some time remained controversial. Further doubt on their validity

analyte + antibody* ⇔ analyte : antibody* (B)
+ residual analyte
+ residual antibody* (F)

Measure "fraction bound" (B) : $[Ab]_{opt} \rightarrow \infty$
Measure "fraction free" (F) : $[Ab]_{opt} \rightarrow 0$

Fig. 3. Labeled-analyte (*left*) and labeled-antibody (*right*) assay systems compared

Labeled-analyte assay systems essentially rely on observation of an analyte "marker" to reveal the products of the reaction between analyte and antibody (although the labeled analyte is not necessarily identical to the unlabeled analyte in its binding characteristics *vis-à-vis* antibody). Note that, irrespective of which fraction of the labeled analyte is measured after the binding reaction, the optimal antibody concentration required to maximize sensitivity in such a system tends toward zero (assuming a background signal of 0). Labeled-antibody systems rely on observation of an antibody "marker" to reveal the products of the binding reaction between analyte and antibody. In this case, the optimal antibody concentration required to maximize sensitivity tends toward zero when the "free" antibody fraction is measured, but tends toward infinity when the bound fraction is determined (likewise assuming zero background)

was cast by the publication by Rodbard and Weiss in 1973 (*14*) of detailed theoretical studies demonstrating that both labeled analyte and labeled antibody methods possessed essentially equal sensitivities. (Note: These authors suggested that IRMAs might be more sensitive in the assay of small polypeptides, in which radioiodine incorporation into the antigen molecule was restricted; conversely, these assays would be less sensitive for the measurement of antigens of high molecular mass.) Nevertheless, despite the appearance of this publication, the belief that labeled antibody methods per se are intrinsically more sensitive than the corresponding labeled analyte methods gained wide acceptance among clinical chemists.

The reason for confusion on this issue is that the greater potential sensitivity of certain assay formats is not really a consequence of the labeling of antibody as opposed to analyte; indeed, the apparent antithesis between labeled-analyte and labeled-antibody methods diverts attention from the true reasons underlying the superior sensitivity of certain assay designs. Theoretical analysis (see, e.g., *4*, *15*) reveals that, assuming "perfect" separation of the products of the binding reaction (i.e., no misclassification of bound and free moieties), the optimal antibody concentration (for maximal sensitivity) in a labeled analyte immunoassay invariably tends to zero, irrespective of whether the free or bound labeled analyte fraction is measured, whereas in labeled-antibody methods the optimal antibody concentration depends on which labeled-antibody fraction is measured (see Figure 3). If the free (unreacted) antibody fraction is measured, the optimal concentration also tends to zero; conversely, if the analyte-bound fraction is measured, the concentration tends to infinity. In short, of the four basic measurement strategies available—labeled analyte, with measurement of free or bound reaction product, and labeled antibody, also with measurement of free or bound product—only one permits, in practice, the use of antibody concentrations approaching infinity.

This particular approach may, for want of a better term, be described as "noncompetitive," although it must be emphasized that such terminology involves a departure from the original meanings attached to "competitive" and "noncompetitive" when these descriptions were first used in the present context. Indeed, as discussed below, assays may be subclassified in this manner when no labeled reagent of any kind is involved.

However, the categorization of immunoassays and other binding assays as competitive or noncompetitive, depending on the binding agent concentration yielding maximal assay sensitivity, itself obscures the underlying reasons for the existence of this divergence in assay designs, and may thus be misleading. These reasons may be more readily understood if the basic principles of such assays are portrayed differently from their customary presentation.

The "Antibody Occupancy Principle" of Immunoassay

When a "sensor" antibody is introduced into an analyte-containing medium, binding sites on the antibody are occupied by analyte molecules to a fractional extent that reflects both the equilibrium constant governing the binding reaction, and the final concentration of free analyte present in the mixture. This proposition stems immediately from the Law of Mass Action, which can be written as

$$[AbAg]/[fAb] = K[fAg] \qquad (1)$$

or as fractional occupancy of antibody binding sites, given by

$$[AbAg]/[Ab] = K[fAg]/(1 + K[fAg]) \qquad (2)$$

where [AbAg], [Ab], [fAb], and [fAg] represent the concentrations (at equilibrium) of bound and total antibody, and free antibody and antigen (analyte), respectively, and K = equilibrium constant. The final concentration of free analyte generally depends on the concentrations of both total analyte and antibody; however, when total antibody approximates $0.05/K$ or less, free and total antigen ([Ag]) concentrations do not differ significantly, and fractional occupancy of antibody is given by

$$[AbAg]/[Ab] = K[Ag]/(1 + K[Ag]) \qquad (3)$$

Assays utilizing this concept have been termed "ambient analyte immunoassays" (*16*), fractional occupancy being independent of both sample volume and antibody concentration (see below).

All immunoassays essentially depend on measurement of the "fractional occupancy" of the sensor antibody after its reaction with analyte (see Figure 4). Techniques relying on the measurement of unoccupied antibody binding sites (from which antibody occupancy is implicitly deduced by subtraction) necessitate—for attainment of maximal sensitivity—the use of sensor antibody concentrations tending to zero; these assays

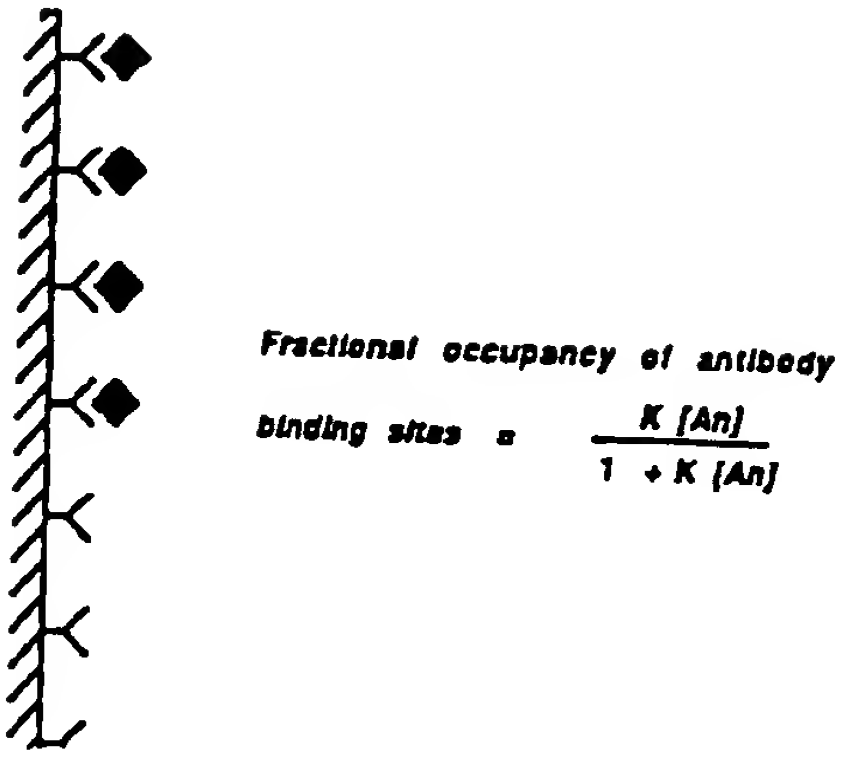

Fractional occupancy of antibody binding sites = $\frac{K\,[An]}{1 + K\,[An]}$

Fig. 4. The antibody binding-site occupancy principle of immunoassay

All immunoassays implicitly rely on the measurement of (fractional) binding-site occupancy by analyte

may therefore be categorized as "competitive." Conversely, techniques in which occupied sites are *directly* measured permit (in principle) the use of relatively high concentrations of sensor antibody and may be described as "noncompetitive." This difference in assay design simply reflects the proposition that, to minimize error in the measurement, it is generally undesirable to measure a small quantity by estimating the difference between two large quantities.

These concepts are illustrated in Figure 5, which portrays basic immunoassay formats currently in common use. Conventional RIA and other similar "labeled-analyte" techniques rely on measurement of *unoccupied* binding sites, generally by back-titration (either simultaneous or sequential) with labeled analyte, but anti-idiotypic antibody (reactive only with unoccupied sites on the sensor antibody) may be used for the same purpose. In the case of single-site labeled-antibody assays, the labeled antibody itself constitutes the sensor antibody; after reaction with analyte, this sensor antibody may be separated into occupied and unoccupied fractions through use of (e.g.) an immunosorbant (comprising antigen, antigen analog, or anti-idiotypic antibody linked to a solid support). If, after separation, the "signal" emitted by labeled antibody *bound* to analyte (i.e., the "occupied" fraction) is measured directly, the assay can be classed as "noncompetitive." Conversely, if one measures the labeled antibody *not bound* to analyte (i.e., that attached to the immunosorbant), then the assay is "competitive."

Two-site "sandwich" assays are clearly more complex because they rely on two antibodies and can be considered from two points of view. For our present purposes, the solid-phase antibody can be regarded as the "sensor" antibody, with the labeled antibody enabling the occupied sensor-antibody binding sites to be distinguished. Seen from this viewpoint, two-site assays may be classed as "noncompetitive."

These considerations emphasize that the differences in design distinguishing so-called competitive and noncompetitive methods are essentially unrelated to which

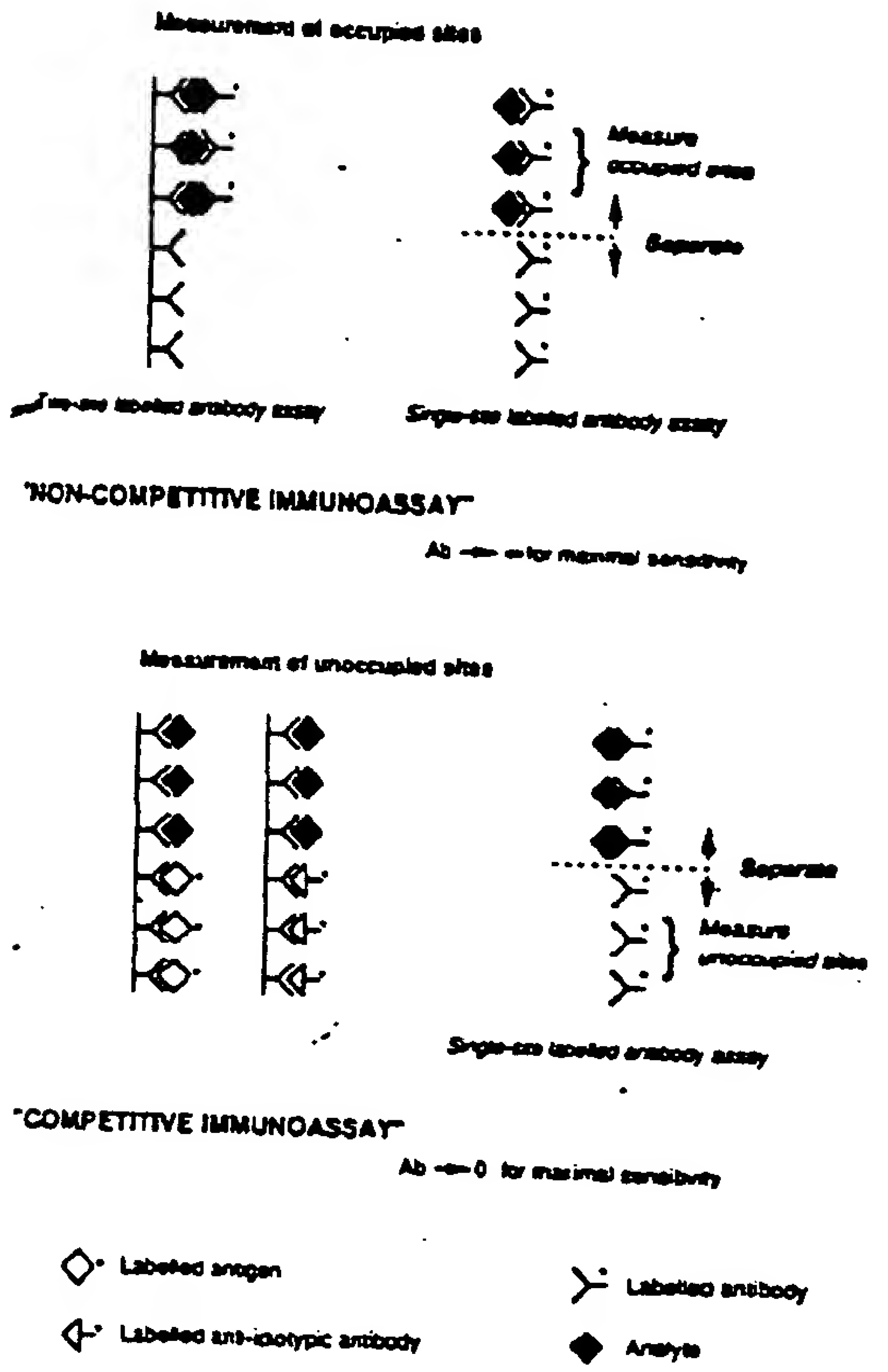

Measurement of occupied sites
Measure occupied sites
Separate
Two-site labeled antibody assay
Single-site labeled antibody assay
"NON-COMPETITIVE IMMUNOASSAY"
Ab → ∞ for maximal sensitivity
Measurement of unoccupied sites
Separate
Measure unoccupied sites
Single-site labeled antibody assay
"COMPETITIVE IMMUNOASSAY"
Ab → 0 for maximal sensitivity
Labeled antigen
Labeled antibody
Labeled anti-idiotypic antibody
Analyte

Fig. 5. Basic competitive and noncompetitive immunoassay designs

The distinction between noncompetitive and competitive immunoassays reflects the way in which antibody binding-site occupancy is observed. Labeled-antibody methods are "noncompetitive" if occupied sites of the (labeled) antibody are directly measured, but are "competitive" (*lower right*) when unoccupied sites are measured. Labeled-antigen (*lower left*) or labeled-anti-idiotypic-antibody methods (*lower center*) rely on measurement of sites unoccupied by analyte, and are therefore of "competitive" design

component (if any) of the reaction system is labeled. Indeed, in the case of transducer-based "immunosensors," no component is labeled; nevertheless, the design of the immunosensor will differ significantly, depending on whether a measurable signal is yielded by occupied or unoccupied antibody binding sites situated on its surface. In short, the terms "competitive" and "noncompetitive" merely reflect alternative approaches to the determination of the occupancy of antibody binding sites and lead to differences in the optimal antibody concentration required to minimize the effects of random errors arising in the determination.

Competitive and noncompetitive immunoassays can be shown to differ significantly in many of their performance characteristics, including their sensitivities. In both types of assays, both the affinity constant (K) of the antibody and the specific activity of the label are important in determining sensitivity; however, in practice, the sensitivity of competitive assays is primarily limited by the affinity constant of the antibody, whereas the specific activity of the label is more important in noncompetitive systems. In both cases, the "experimental" or "manipulation" error in the measurement of the zero-dose response (R_0) [i.e., the relative error (σ_{R_0}/R_0) arising from pipetting and other operations, but not including the statistical signal measurement error se] is of key importance in determining "potential" assay sensitivity (i.e., the sensitivity obtained by assuming the specific activity of the label to be infinite, implying zero error in signal measurement). Thus the potential sensitivity of a competitive assay can be shown to be σ_{R_0}/KR_0, whereas that of a noncompetitive assay is given by $R_0\sigma_{R_0}/[\mathrm{Ab}]KR_0$, where, in the latter case, R_0 is assumed to represent the labeled antibody misclassified as bound ($[\mathrm{bAb}]_0$), commonly referred to as "nonspecifically bound" antibody. Thus $R_0/[\mathrm{Ab}] = f$, the fraction of labeled antibody that is nonspecifically bound, and $R_0\sigma_{R_0}/[\mathrm{Ab}]KR_0 = f\sigma_{R_0}/KR_0$. Assuming that the relative error (σ_{R_0}/R_0) in the measurement of the zero-dose response is approximately identical for both competitive and noncompetitive assays, it is evident from this simple analysis that the potential sensitivity of noncompetitive methods is greater than that of competitive methods by the factor f, i.e., by the fraction of labeled antibody that is "nonspecifically bound." For example, if the nonspecifically bound fraction is 0.01%, a noncompetitive strategy is potentially capable of a sensitivity 10 000-fold greater than that of a competitive approach, other factors being equal.

These findings are summarized in Figure 6 (left), which shows the relationships between sensitivity (expressed in terms of molecules per milliliter) and antibody affinity in an optimized competitive (labeled analyte) assay. For this analysis, we assume (*a*) the use of a label of infinite specific activity, and (*b*) the use of ^{125}I as a label, the radioactivity of the samples being counted for 1 min. Computations of the theoretically optimal reagent concentrations (on which calculations represented in Figure 6 rely) were based on the further assumptions that (*c*) the radioactivity of the antibody-bound labeled-analyte fraction was counted and (*d*) the (relative) "experimental error" component in the measurement of the bound fraction (σ_b/b) was 1%. Given these assumptions, the "potential" sensitivity attainable in such an assay is σ_b/Kb, where *K* is the affinity constant of the antibody. [For example, if the affinity constant is 10^{12} L/mol, and σ_b/b is 0.01 (1%), maximal assay sensitivity is 10^{-14} mol/L, or $\sim 6 \times 10^6$ molecules/mL.] The additional "signal measurement error" arising in consequence of counting radioactive samples for a finite time implies a loss of assay sensitivity, as shown by the upper curve in Figure 6 (left). However, the resulting loss in sensitivity is relatively small for antibodies of affinities $<10^{12}$ L/mol, and is negligible for antibodies with affinities $<10^{11}$ L/mol. In other words, if the assayist can accept individual sample counting times of 1–5 min, little improvement in sensitivity is gained by using alternative labels of higher specific activities than ^{125}I. However, similar considerations suggest that radioisotopic labels of much lower specific activity than ^{125}I (e.g., ^{3}H) may limit the sensitivities of the assays (such as steroid assays) in which they are used, notwithstanding the use of relatively long sample counting times.

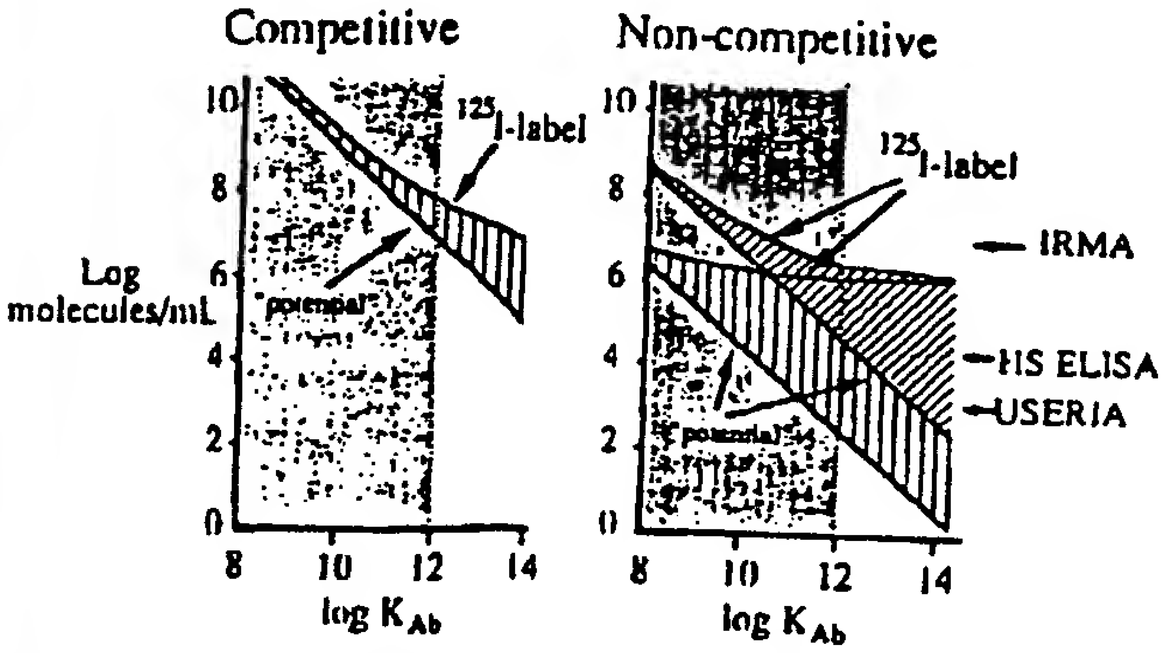


Fig. 6. Theoretically predicted sensitivities of competitive and noncompetitive immunoassay methods (represented by the SD of zero analyte measurements, expressed as molecules/mL) plotted as a function of antibody affinity (*K*)

Note: in noncompetitive sandwich assays, the antibody affinity referred to is that of the labeled antibody. In the competitive assays, calculations are based on the assumption that the experimental error (CV) incurred in the measurement of the assay response (*e.g.*, fraction of labeled antigen bound) is 1%. The "potential sensitivity" curve assumes the use of a label of infinite specific activity, implying that the error in the measurement of the label per se is zero. The *^{125}I-label curve* indicates the loss in sensitivity arising from the statistical error incurred in counting ^{125}I disintegrations for a finite counting time. Note that, if using antibodies with an affinity $<10^{12}$ L/mol (the maximum achieved in practice), little increase in sensitivity can be achieved by using labels of higher specific activity than ^{125}I. For noncompetitive assays, the potential sensitivity curves shown relate to values of nonspecific binding of labeled antibody of 1% (*upper curves*) and 0.01% (*lower curves*), and emphasize the improvement in sensitivity potentially attainable by minimizing nonspecific binding. The corresponding *^{125}I-label curves* demonstrate the much greater loss in sensitivity (compared with that potentially attainable) when a radioisotopic marker is used, and the special advantages of nonisotopic labels of higher specific activity in noncompetitive assay designs (particularly if nonspecific binding is reduced to 0.1% or less). *Arrows* indicate assay sensitivities reported for noncompetitive immunoassays based on ^{125}I (IRMA), and enzymes relying on fluorogenic (HS-ELISA) (*28*) and radioactive (USERIA) (*29*) substrates. These conclusions underlay the original development (*19, 20*) of time-resolved fluoroimmunoassay (DELFIA), the first nonisotopic "ultra-sensitive" immunoassay methodology.

The other main conclusions stemming from such analysis are the importance of both minimizing "manipulation" errors and using antibodies of high binding affinity. For example, an increase in σ_b/b to 3% implies an approximate threefold loss in sensitivity, notwithstanding the fact that an assay reoptimized in response to the deterioration in operator skill that these numbers imply would utilize less antibody and labeled analyte, thereby partially offsetting the consequences of poor pipetting. But the most important conclusion emerging from the analysis is the near impossibility, in practice, of achieving immunoassay sensitivities better than about 10^7 molecules/mL by using a competitive approach, irrespective of the nature of the label used, if one assumes an upper limit to antibody binding affinities on the order of 10^{12} L/mol.

The results of a similar analysis of the sensitivity limitations applying to noncompetitive (two-site) assays (*15*) are illustrated in Figure 6 (right). Two sets of curves are portrayed here, corresponding to the assumptions of 1% and 0.01% nonspecific binding of labeled antibody to the capture-antibody substrate. Such analysis likewise yields important conclusions relevant to assay design, e.g., the crucial importance of reducing nonspecific binding of labeled antibody to an absolute minimum. Furthermore, if nonspecific binding is reduced to ~0.01%, just as high sensitivity is achievable by using an antibody of $K = 10^8$ L/mol in an optimized noncompetitive assay design as by using an antibody of $K = 10^{12}$ L/mol in a competitive method. One of the most important conclusions is that the sensitivities potentially attainable with high-affinity antibodies ($K > 10^{10}$ L/mol) are beyond the reach of radioisotopically based methods, which (because of the relatively low specific activities of isotopes such as ^{125}I) are limited in practice to sensitivities of the order of 10^6–10^7 molecules/mL or more. In short, although, under certain circumstances, noncompetitive IRMAs may be somewhat more sensitive than corresponding RIA techniques (assuming the use of the same antibody in each methodology), the potential advantages (*vis-à-vis* sensitivity) of the noncompetitive approach can be realized only by using nonisotopic labels of much higher specific activity than ^{125}I. The superiority of such labels is most apparent when they are combined with high-affinity antibodies; however, Figure 6 demonstrates that, even with use of antibodies with affinities of about 10^8–10^9 L/mol, nonisotopic labels may yield a substantial improvement in sensitivity.

These theoretical conclusions, together with the publication by Köhler and Milstein (*18*) of methods of in vitro production of monoclonal antibodies (*1*), constituted the basis of my laboratory's collaborative development (initiated around 1976) with the instrument manufacturer LKB/Wallac of the time-resolved fluorometric immunoassay methodology now known as DELFIA (*19, 20*). This methodology was the first "ultra-sensitive" nonisotopic immunoassay methodology to be developed. The same basic approach has subsequently been adopted by many other manufacturers, using a variety of high-specific activity labels (Table 1).

Against this background, let us now turn to the development of highly sensitive, miniaturized "microspot" immunoassays and multianalyte assay systems.

Antibody "Microspot" Immunoassay: Basic Concepts and Theory

Ambient Analyte Immunoassay

Particular attention has been drawn above to the specious notion that an antibody concentration approximating $0.5/K$ is required to maximize the sensitivity of conventional labeled-antigen assays. This proposition is implicitly overturned by the development of "microspot" immunoassays, which we expect to provide the basis of a new generation of binding assay methods. But before discussing this methodology in detail, another basic analytical concept must be examined.

Table 1. Detection Limits According to Type of Label

Label	Specific activity
^{125}I	1 detectable event per second per 7.5 × 10^6 labeled molecules
Enzyme label	Determined by enzyme "amplification factor" and detectability of reaction product
Chemiluminescent label	1 detectable event per labeled molecule
Fluorescent label	Many detectable events per labeled molecule

The recognition that all immunoassays essentially rely on measurement of antibody occupancy leads to a potentially important type of assay, ambient analyte immunoassay (*16*). This name is intended to describe assay systems that, unlike conventional methods, measure the analyte *concentration* in the medium to which an antibody is exposed, being independent both of sample volume and of the amount of antibody present. The possibility of developing such assays follows from the Law of Mass Action, which leads to the following equation, representing the fractional occupancy (F) by analyte of antibody binding sites (at equilibrium):

$$F^2 - F\{(1/[\underline{Ab}]) + ([\underline{An}]/[\underline{Ab}]) + 1\} + [\underline{An}]/[\underline{Ab}] = 0 \qquad (4)$$

where $[\underline{An}]$ = analyte concentration, $[\underline{Ab}]$ = antibody concentration (both in units of $1/K$).[1]

From this equation it may readily be shown that, for antibody concentrations approaching 0, $F = [\underline{An}]/(1 + [\underline{An}])$. This conclusion is illustrated in Figure 7, in which the fractional occupancy of ("monospecific" or "monoclonal") antibody binding sites in the presence of various analyte concentrations is plotted against antibody concentration. When an antibody concentration of less than (say) $0.01/K$ (the antibody preferably, but not essentially, being coupled to a solid support) is exposed to an analyte-containing medium, the resulting (fractional) occupancy of antibody binding sites solely reflects the ambient concentration of analyte[1] and is independent of the total amount of antibody in the system. (If, for example, $K = 10^{11}$ L/mol, an antibody binding-site concentration of $0.01/K$ represents 0.01×10^{-11} mol/L, or 6.02×10^7 binding sites/mL.) Analyte binding by antibody causes depletion of (unbound) analyte in the medium but, because the amount bound is small, the resulting reduction in the ambient concentration of analyte is insignificant. For example, if the concentration of binding sites of the sensor antibodies is $<0.01/K$, analyte depletion in the medium is invariably <1%, and the system is therefore effectively indepen-

[1] Expression of reagent concentrations in terms of $1/K$ units has the effect of generalizing the graphical representation of binding assay data. The terms $[\underline{Ab}]$ and $[\underline{An}]$ are underlined to indicate that this convention has been adhered to in deriving equation 4. They do not refer to molar concentrations and are not interchangeable with [Ab] and [An]. For example, if the antibody possesses an affinity (constant) for analyte of 10^{11} L/mol, a concentration of 10^{-11} mol/L (represented in units of $1/K$) is 1 (dimensionless) unit. Thus, fractional occupancy curves based on equation 4 are identical for *all* antibodies if this way of expressing antibody concentration is adopted: i.e., curves relating F to analyte concentration will be identical for systems using 10^{-11} mol/L concentrations of an antibody with an affinity of 10^{11} L/mol, 10^{-10} mol/L of an antibody with an affinity of 10^{10} L/mol, 10^{-9} mol/L of an antibody with an affinity of 10^9 L/mol, etc. (provided the analyte concentration is expressed in the same manner).

[2] The term "ambient" is used to indicate that antibody occupancy reflects the analyte concentration to which antibody binding sites are exposed, not the amount of analyte in the incubation tube; i.e., the system is independent of sample volume

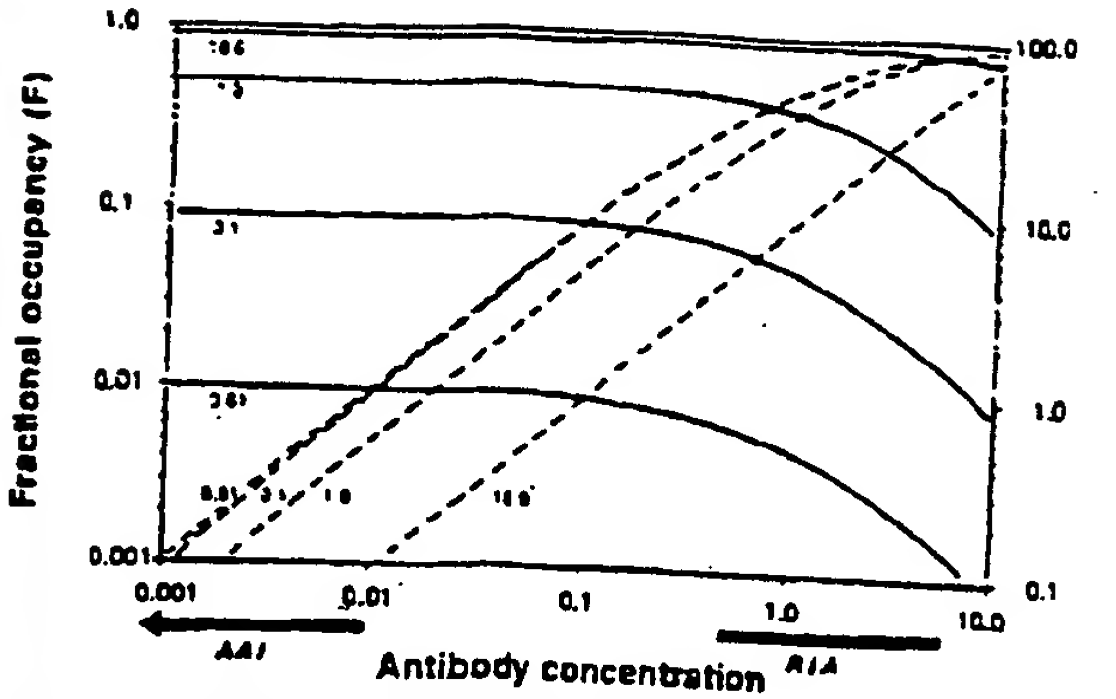


Fig. 7. Fractional antibody binding-site occupancy (F, see equation 4) plotted as a function of antibody binding-site concentration for different values of analyte (antigen) concentration (—), and the percentage binding (b) of analyte to antibody (*right-hand ordinate*; – – –)

All concentrations are expressed in units of 1/*K*. Note that for antibody concentrations <0.01/*K* (approximately), the percentage binding of analyte is <1% for all analyte concentrations, and fractional binding-site occupancy is essentially unaffected by variations in antibody concentration extending over several orders of magnitude, being governed solely by antigen concentration (ambient analyte immunoassay). Note that radioimmunoassays and other "competitive" immunoassays are conventionally designed to use antibody concentrations approximating 0.5/*K*–1/*K* or more (implying binding of analyte concentrations tending to zero (b_0) >30%), in accordance with the precepts of Yalow and Berson (*1, 3*)

dent of sample volume.

These conclusions lead to two further concepts. First, the antibody may be confined to a "microspot" on a solid support, such that the total number of antibody binding sites within the microspot is $<v/K \times 10^{-5} \times N$, where v = the sample volume to which the microspot is exposed (in milliliters) and N = Avogadro's number (6×10^{23}). For example, if $v = 1$ and $K = 10^{12}$ L/mol, then the maximum number of binding sites that will cause negligible disturbance (<1%) to the ambient concentration of analyte is 6×10^6, this number being greater for lower-affinity antibodies. Furthermore, the perception that the ratio of occupied (or unoccupied) sites to total binding sites is solely dependent on the ambient concentration of analyte leads to the concept of a dual-label, "ratiometric," microspot immunoassay.

Dual-Label Microspot Immunoassay

After exposure of a microspot of antibody (located on a suitable probe) to an analyte-containing fluid (see Figure 8, left), the probe may be removed and exposed to a solution containing a high concentration of a "developing" antibody directed against either a second epitope (i.e., the occupied site) on the analyte molecule if the molecule is large, or against unoccupied binding sites on the antibody in the case of small analyte molecules (Figure 8, right). The fractional occupancy of the sensor antibody may thus be estimated by measuring the ratio of sensor and developing antibodies that form the dual-antibody "couplets." This can be readily achieved by labeling the sensor and the developing antibodies with different labels, e.g., a pair of radioactive, enzyme, or chemiluminescent markers (or even labels of entirely different nature). Fluorescent labels are potentially particularly useful in this context because, by the use of optical scanning techniques (Figure 9), they permit the scanning of arrays of antibody "microspots" distributed over a surface (each microspot directed against a different analyte), so that multiple analyte assays may be performed simultaneously on the same sample. Several

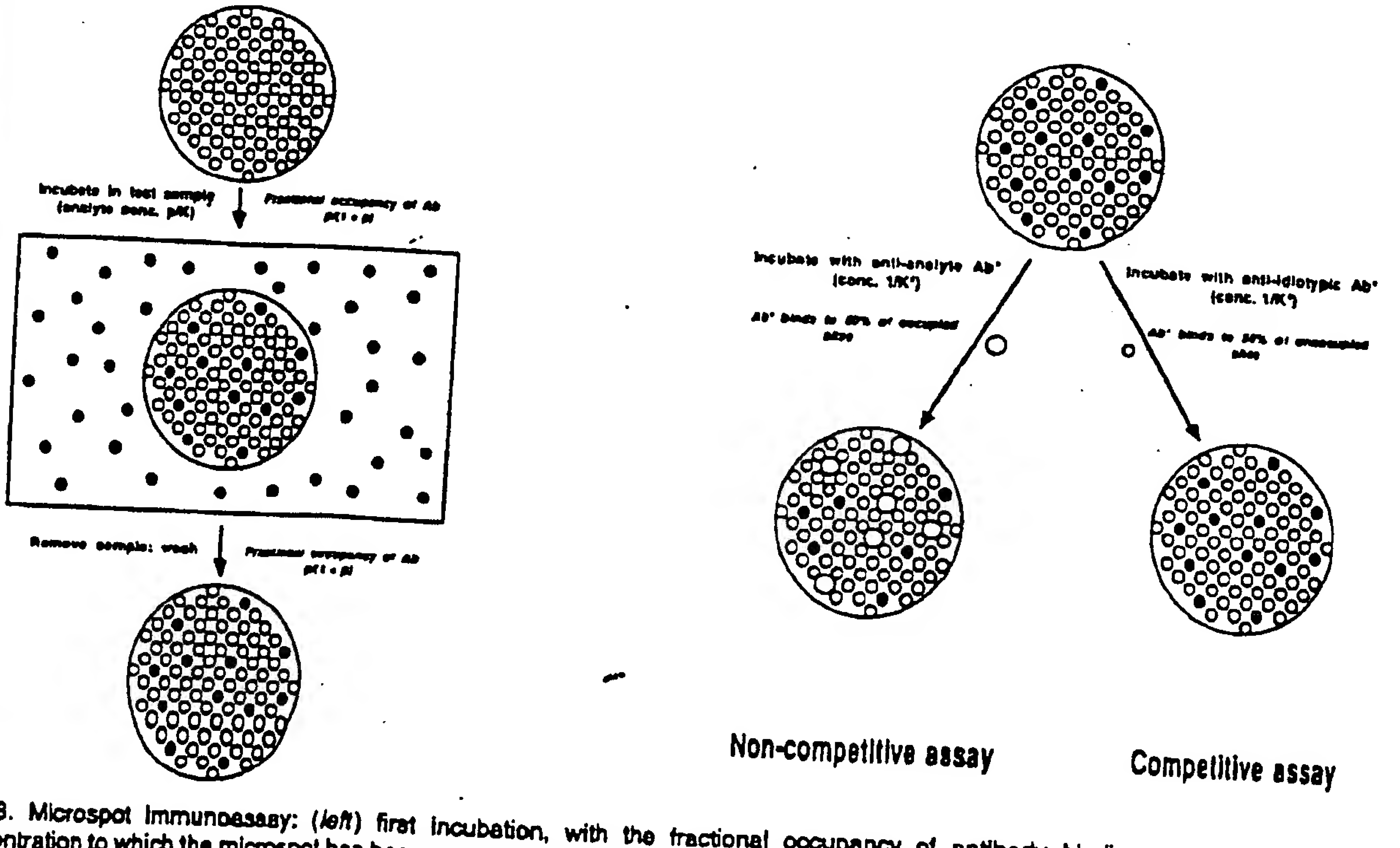

Incubate in test sample (analyte conc. p/K)
Fractional occupancy of Ab p/(1 + p)
Remove sample: wash
Fractional occupancy of Ab p/(1 + p)
Incubate with anti-analyte Ab* (conc. 1/K*)
Ab* binds to 50% of occupied sites
Incubate with anti-idiotypic Ab* (conc. 1/K*)
Ab* binds to 50% of unoccupied sites
Non-competitive assay
Competitive assay

Fig. 8. Microspot immunoassay: (*left*) first incubation, with the fractional occupancy of antibody binding sites reflecting the analyte concentration to which the microspot has been exposed; (*right*) second incubation, in which the microspot is exposed to a second "developing" antibody reactive with either occupied sites (noncompetitive assay), or unoccupied sites (competitive assay)

In the second incubation, a concentration of developing antibody has been selected such that only 50% of the occupied or unoccupied sites is identified

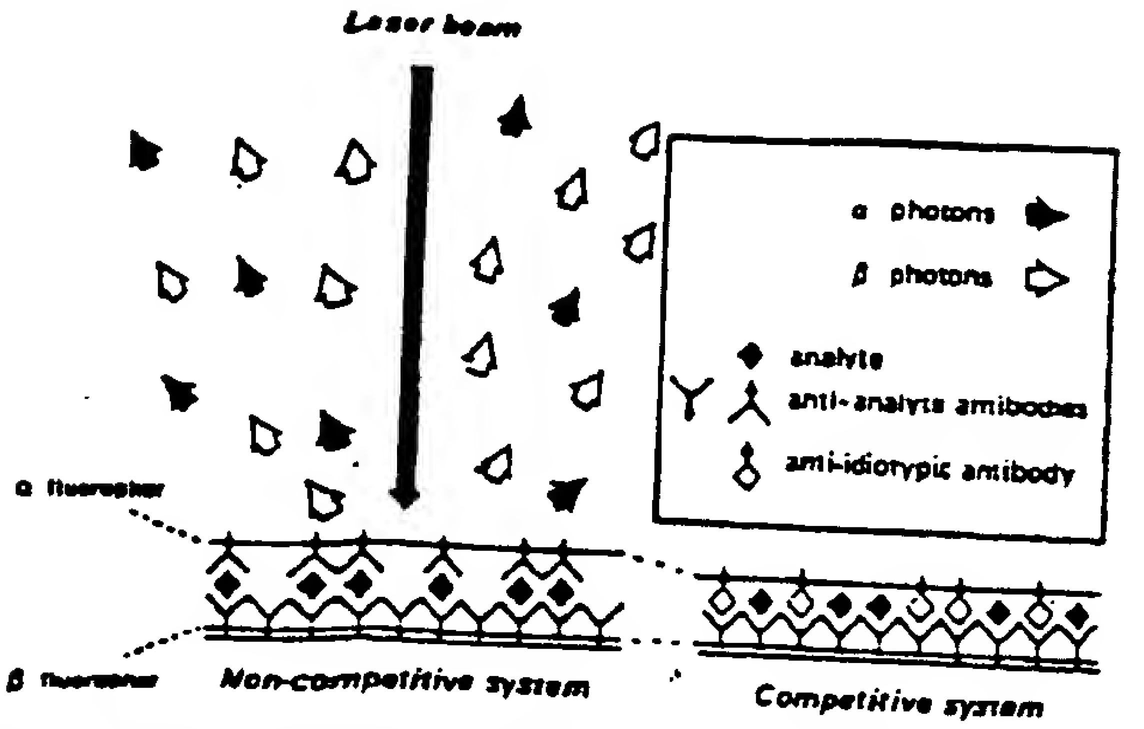


Fig. 9. Basic principle of dual-label, ambient analyte immunoassay relying on fluorescent-labeled antibodies

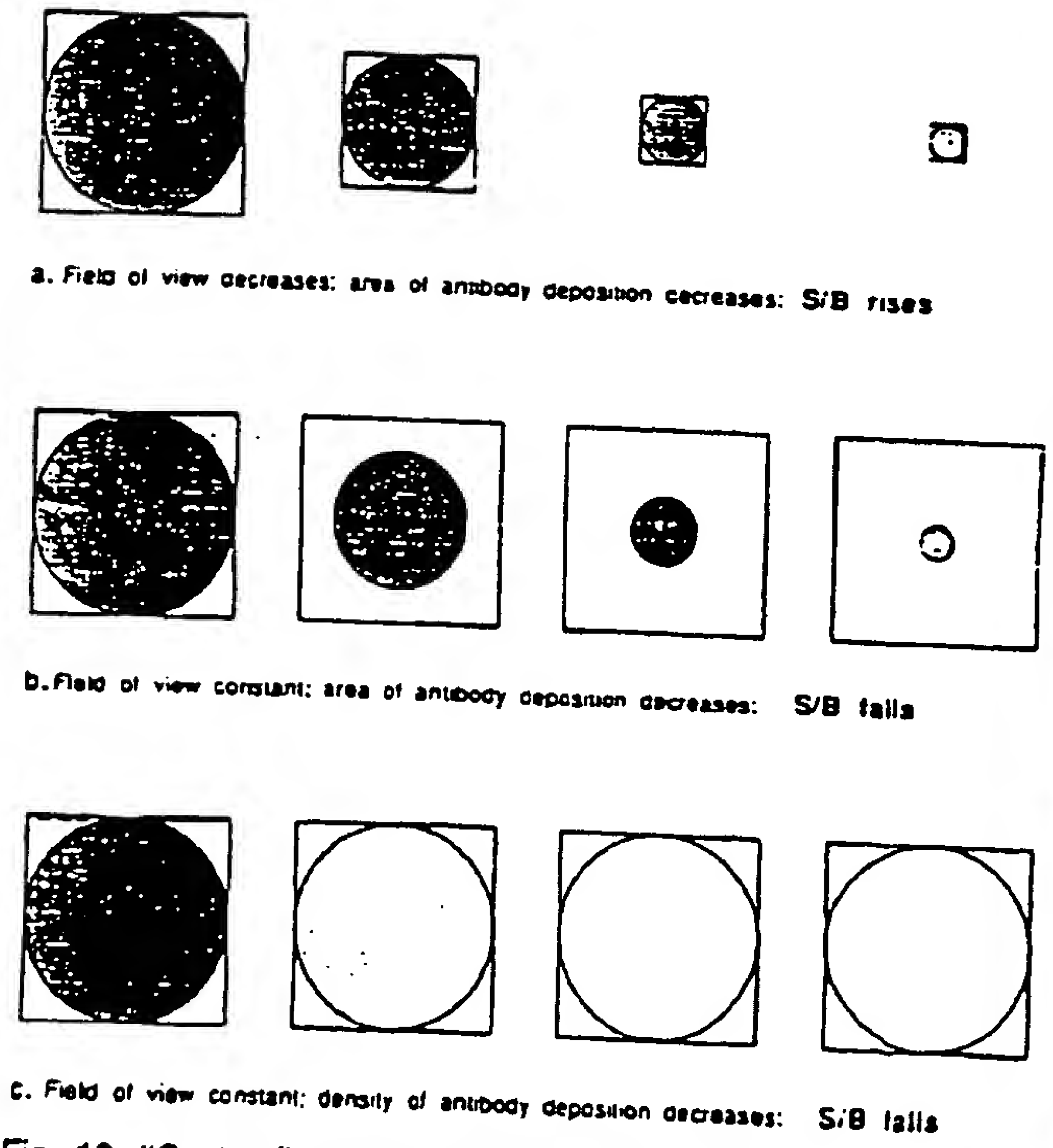
The ratio of α and β fluorescent photons emitted reflects the value of F (see Fig. 7) and depends solely on the analyte concentration to which the probe has been exposed. The ratio is unaffected by the amount or distribution of antibody coated (as a monomolecular layer) onto the probe surface

advantages stem from adopting a dual fluorescence measurement. For example, neither the amount nor the distribution of the sensor antibody within the detector's field of view is important, because the ratio of the emitted fluorescent signals is unaffected. Likewise, fluctuations in the intensity of the incident (exciting) light beam are apt to be of little significance. These advantages are additional to the basic benefit stemming from this approach, i.e., that the necessity of ensuring constancy of the amount of sensor antibody used in the assay system is removed.

Microspot Immunoassay Sensitivity

Because the microspot immunoassay methodology challenges concepts that have dominated immunoassay design theory in the past two to three decades, consideration of the potential sensitivity attainable by this approach is obviously of primary importance. The proposition that microspot assays may be at least as sensitive as conventional systems that rely on far larger amounts of antibody may readily be demonstrated by consideration of a model system. Let us postulate that sensor antibody molecules are attached to the surface of a solid support such that their binding sites remain exposed to the analyte, and that their affinity for the analyte is thereby unchanged. (The antibody concentration in the system—the number of binding sites on the support divided by the incubation volume—is unaffected by such attachment, and antibody occupancy by analyte at equilibrium will be identical to that occurring if the antibody is distributed uniformly throughout the incubation mixture.) Let us also suppose that the antibody molecules exist as a uniform monolayer of maximal surface density on the support and (to simplify discussion) are unlabeled. Then a change in the concentration of sensor antibody implies a corresponding change in the surface area over which the antibody is distributed. If, for example, the antibody affinity constant is 10^{11} L/mol, the total incubation volume is 1 mL, and the antibody surface density is 6000 binding sites/μm^2, then a surface area of 10^5 μm^2 (i.e., 0.1 mm^2) accommodates antibody binding sites corresponding to a concentration of 0.1/*K*; an area of 0.01 mm^2 corresponds to a concentration of 0.01/*K*, etc. Let us further postulate that, after exposure of the sensor antibodies to a medium containing analyte at a concentration of 0.01/*K* (i.e., 6 × 10^7 molecules/mL), we measure "noncompetitively" the resulting antibody occupancy (e.g., by exposure to a second, labeled, "developing" antibody directed against the analyte, forming a typical antibody sandwich). Finally, let us suppose that all occupied sites react with the developing antibody, with the latter also binding "nonspecifically" to the solid support itself at a surface density of 1 molecule/μm^2.

We may now consider the effects of a progressive reduction of the antibody-coated surface area from (e.g.) 1 mm^2 (effective antibody concentration 1/*K*) through 0.1 mm^2 (0.1/*K*) to 0.01 mm^2 (0.01/*K*) and below. From equation 4, the value of F for the 1 mm^2 area is 4.98 × 10^{-3}. Thus at equilibrium the number of analyte and labeled antibody molecules specifically bound to the area is 2.99 × 10^7 (i.e., about 50% of the total analyte molecules present), whereas the number of labeled antibody molecules nonspecifically bound is 10^6. Thus, assuming the field of view of the detecting instrument is restricted to the area on which the sensor antibody is deposited (see Figure 10*a*), and (provisionally) assuming the background (or "noise") of the instrument itself to be zero (i.e., the only source of background is the non-


a. Field of view decreases: area of antibody deposition decreases: S/B rises

b. Field of view constant: area of antibody deposition decreases: S/B falls

c. Field of view constant: density of antibody deposition decreases: S/B falls

Fig. 10. "Capture" antibody (CAb) is assumed coated on circular (*shaded*) areas; the field of view of the signal-measuring instrument is represented by square (*unshaded*) areas

(*a*) Reduction of both the area of deposition of CAb and the field of view results in an increase in the signal/noise (S/B) ratio. If the CAb is reduced either by reducing the antibody coated area (*b*) or the density of antibody coating (*c*) while the field of view remains unchanged, S/B falls specifically-bound labeled antibody within the instrument's field of view), the signal/noise ratio observed for the 1 mm^2 area is ~30. Similarly, the value of F for a 0.1 mm^2 area is 9.02×10^{-3}, the number of labeled antibody molecules specifically bound to the area is 5.41×10^6, the number nonspecifically bound is 10^6, and the signal/noise ratio is ~54. Likewise, the signal/noise ratio for a 0.01 mm^2 area can be shown to be ~59. In short, the signal/noise ratio increases as the antibody-coated surface area is decreased, approaching a maximal (plateau) value of 60 as the area coated with sensor antibody falls below 0.01 mm^2 and tends toward zero.

If, however, a reduction in the antibody-coated area were *not* accompanied by a corresponding reduction in the detecting instrument's field of view, the resulting reduction in "signal" would *not* lead to a corresponding decrease in the background generated by nonspecifically-bound developing antibody (Figure 10*b*). Therefore, although reduction in the coated area would increase the fractional occupancy of the sensor antibody, the signal/noise ratio might either remain constant or fall. In these circumstances it might be advantageous to *increase* the coated area. Similarly, if the surface density of sensor antibody were decreased (the coated area being held constant), similar conclusions would be reached (Figure 10*c*).

Likewise, if the background signal generated within the detecting instrument itself (e.g., from the photocathode of a photomultiplier tube used to detect photons emitted from the antibody-coated area) were not zero, and remained constant regardless of the instrument's field of view, then a maximum signal/noise ratio would also be attained at some optimal value of the antibody-coated area, below which the ratio would fall. Because, however, one can generally reduce the size of the detector (and hence the detector-generated background) at the same rate as the size of the signal-emitting area, there is no reason—in principle—for the signal/noise ratio to diminish as the antibody-coated area is progressively reduced toward zero. Thus if we accept the signal/noise ratio as indicative of the precision of the measurement of antibody occupancy (and hence of assay sensitivity), these considerations suggest that it is advantageous to reduce the antibody-coated surface area (and, concomitantly, the sensor-antibody concentration) toward zero, although little advantage is likely to accrue from reducing the area below 0.01 mm^2 (and thus the antibody concentration below $0.01/K$).

Were the microspot area indeed reduced to zero, both signal and noise would likewise also fall to zero (the ratio between them nevertheless remaining essentially constant), implying that no signal of any kind would, in the limit, be recorded. In practice, other statistical factors come into play when the number of individual events (e.g., photons) observed by a detecting instrument is very low, thus prohibiting a reduction of the sensor antibody concentration to zero. The point at which the reduction in the antibody-coated area causes the detectable signal to be lost sufficiently to affect the precision of the measurement of antibody occupancy depends clearly on the specific activity of the labeled antibody used to measure the occupied binding sites: the higher the specific activity, the smaller the permissible area. Thus, given labels of very high specific activity, one can envision circumstances in which, even in a "noncompetitive" system, the optimal concentration of sensor antibody may be exceedingly low. A more general conclusion is that a variety of factors, including the characteristics of the instruments used for measuring the labeled antibody (or labeled analyte), influence immunoassay design, implying, among other things, the virtual impossibility of formulating general rules regarding this. For example, reagent concentrations that are optimal for isotopically labeled reagents used with a conventional radioisotope counter (possessing a fixed background dependent on its basic construction) are likely to be entirely different when very high-specific-activity labels are used and one has the freedom to tailor the measuring instrument to samples of any size. In short, certain conclusions based on experience of RIA and IRMA techniques may prove misleading when applied to nonisotopic methodologies, and should be viewed with caution.

A more detailed theoretical consideration of (noncompetitive) microspot immunoassay sensitivity *(21)* suggests that

$$C_{min} = D^*_{min} \times [(6 \times 10^{20})(1 + [Ab^*])]/DK[Ab^*] \quad (5)$$

where D = surface density (binding sites/μm^2) of sensor antibody, K = sensor antibody affinity (L/mol), $[Ab^*]$ = concentration of labeled antibody in developing solution (expressed in units of $1/K^*$, where K^* = labeled antibody affinity), D^*_{min} = minimum detectable surface density of labeled antibody (molecules/μm^2), and C_{min} = assay detection limit (molecules/mL). For example, if $[Ab^*] = 1$, $D = 10^5$ molecules/μm^2, $K = 10^{11}$ L/mol, and D^*_{min} = 20 molecules/μm^2, then $C_{min} = 2.4 \times 10^6$ molecules/mL $= 4 \times 10^{-15}$ mol/L and the fractional occupancy of the binding sites of the sensor antibody by the minimum detectable concentration of analyte is 0.04%. Figure 11 shows the theoretical assay sensitivities attainable with use of sensor antibodies of various affinities, plotted as a function of D^*_{min}.

A similar theoretical analysis of competitive microspot immunoassay indicates that potential sensitivities are essentially identical to those attainable with conventional competitive methodologies. In summary, the above considerations indicate that the attainment of high microspot assay sensitivity requires close packing of molecules of sensor antibodies within the microspot area, combined with the use of an instrument capable of accurately measuring very low surface densities of developing antibodies. They also suggest that *(a)* microspot assay sensitivities considerably higher than those obtainable by conventional isotopically based immunoassays are achievable, and *(b)* if labels of very high specific activity are available, the sensitivities yielded

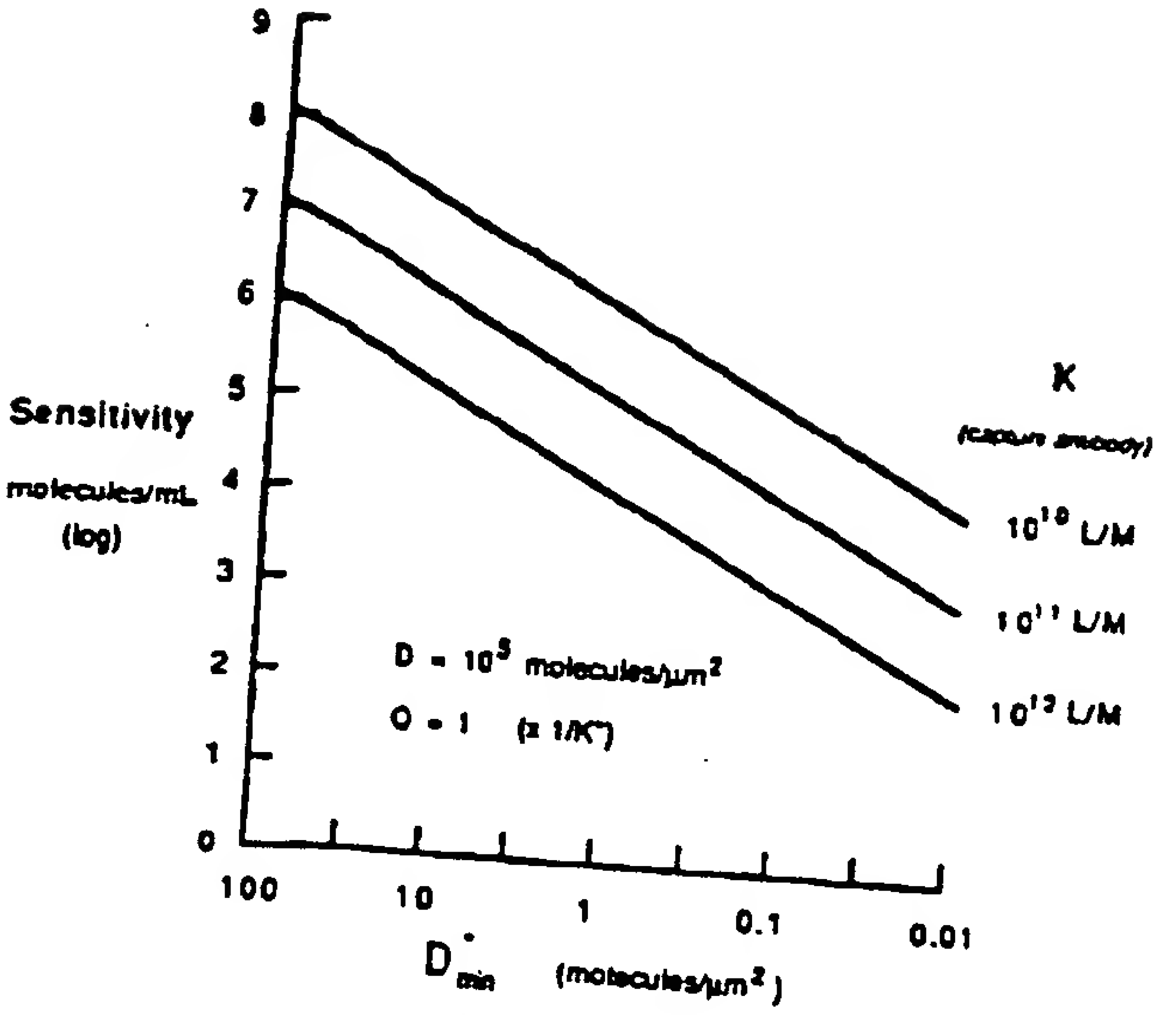


Fig. 11. Theoretically predicted sensitivity of noncompetitive microspot immunoassay plotted as a function of the minimum developing antibody density detectable within the microspot area

Postulated values of capture antibody surface density are 10^8 molecules/μm², and of developing antibody concentration are 1/K Currently available instruments permit detection of between 10 and 1 molecules of fluorescein-labeled antibody per micrometer²

by microspot assays are unlikely to be inferior and (depending on the characteristics of the measuring instruments used) could be superior to the sensitivities achievable in macroscopic assays of conventional design.

Finally, we briefly address a further question occasionally raised in this context, i.e., the kinetic characteristics of microspot assays. Two points should be made regarding this issue. First, the smaller the microspot of sensing antibody, the lower the diffusion constraints on the velocity of the antibody/analyte binding reaction, so that at the limit (i.e., when the amount of antibody situated within the microspot area approaches zero) the kinetics of the reaction approximate those observed in a homogeneous liquid-phase system. Second, although the effective concentration of sensor antibody in the incubation medium is exceedingly low, the fractional rate at which sensor antibody binding sites within the microspot become occupied is invariably greater in this circumstance than when a relatively high concentration of antibody is used, as in conventional assays, particularly those of noncompetitive design. In other words, bearing in mind the relationship between fractional occupancy of sensor antibody and the signal/noise ratio discussed above, it is readily demonstrable that the rate at which the ratio rises is greatest when the microspot area (and the antibody contained within it) is least. Thus, given instrumentation whose field of view is restricted to the microspot area, the highest signal/noise ratio will be observed (after any selected incubation period) when the concentration of sensor antibody in the system is <0.01/*K*. In short, contrary perhaps to superficial impression, and to the generally accepted belief that short immunoassay incubation times require the use of very large amounts of antibody, the antibody microspot approach provides the basis of assays potentially more rapid than any currently available.

Microspot Immunoassay: Some Practical Considerations

Although various high-specific-activity antibody labels are potentially usable in this context, our preliminary studies have relied on the use of conventional fluorophors. The simultaneous measurement of dual fluorescences from small areas is, of course, well established, and the availability of improved instrumentation (e.g., the laser scanning confocal microscope), albeit not specifically designed for the present purpose, has been useful in demonstrating the feasibility of the microspot approach.

In laser scanning confocal fluorescence microscopes, a small area of the specimen is illuminated by a focused laser beam, the fluorescence photons emitted from this area being focused in turn onto a detector, typically a low-dark-current photomultiplier (*22, 23*). At the "confocal" point, the projection of the illumination pinhole and the back-projection of the detector pinhole coincide (Figure 12). Fluorescence photons emitted at other points thus possess a low probability of reaching the detector. Such systems contrast with conventional epifluorescence microscopes, in which the specimen is exposed to an essentially uniform flux of illumination, and yield much sharper images of fluorescent emitters situated in a defined plane of a tissue sample. Electrons spontaneously emitted by the photomultiplier photocathode contribute to the background signal of the instrument, and must—for highest microspot assay sensitivity—be minimized. Fortunately, the design of such instruments permits the photocathode to be very small in area, and this source of background can be expected

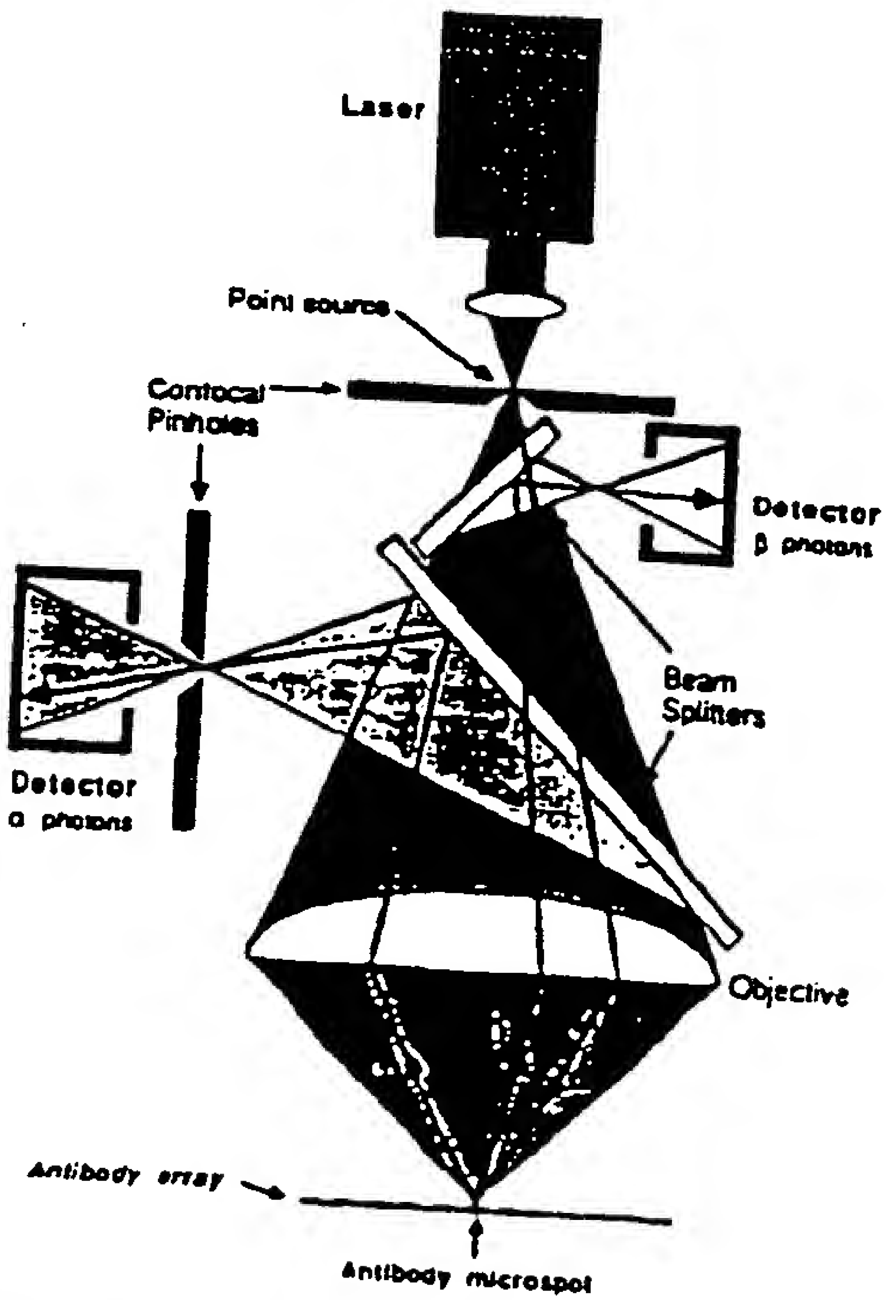


Fig. 12. Schematic diagram of a dual

to diminish with future improvement in photomultiplier design. Other sources of background include fluorescence emitted by components in the optical system, which may not, in current instruments, have been constructed with background reduction as a prime consideration. Nevertheless, they detect with high sensitivity fluorescent signals. For example, one commercially available microscope is claimed to detect fluorescein at a density of 10 molecules/μm^2. Most commercially available fluorescein isothiocyanate (FITC)-labeled IgG exhibits a fluorophor/protein ratio of ~4; this implies detection limit (D^*_{min}) for antibody surface density of two or three FITC-labeled IgG molecules per micrometer^2. This, in turn, implies a theoretical sensitivity for a two-site immunoassay of ~2–3 × 10^5 analyte molecules per milliliter, assuming identical parameter values as above, or 2–3 × 10^4 molecules/mL if the sensing antibody has an affinity of 10^{12} L/mol. Clearly, sensitivity may be increased by loading more fluorophor either directly or indirectly onto the antibody.

Our preliminary studies have relied on a less sensitive microscope, albeit one possessing facilities for dual-fluorescence measurement. Its argon laser emits two excitation lines at 488 and 514 nm. It is thus particularly efficient in exciting blue/green-emitting fluorophores such as FITC (excitation maximum 492 nm), but is less efficient in exciting fluorophores such as Texas Red (excitation maximum 596 nm). However, the ratiometric assay principle permits considerable variation in detection efficiencies of the two labels because the specific activities of the labeled antibody species forming the antibody couplets can be chosen to yield signal ratios approximating unity. Inefficiency of the argon laser in exciting Texas Red is thus not a major handicap in this context. Though this instrument relies on a conventional microscope and not on an optical system designed for this purpose (and thus implicitly less sensitive), it permits quantification of fluorescence signals generated from microspots of any selected area. Initial studies have revealed that, under conditions that are not optimal, the instrument is capable of detecting ~25 FITC-labeled and (or) 150 Texas Red-labeled IgG molecules per micrometer^2, while scanning an area of ~50 μm^2.

The development of microspot immunoassays has also necessitated closer scrutiny of the mechanisms involved in the coupling of antibodies to solid supports. In the present context, these should display a capacity to adsorb (in the form of a monolayer)—or to covalently link—a high surface density of antibody combined with low intrinsic-signal-generating properties (e.g., low intrinsic fluorescence), thus minimizing background. We have examined a number of candidate materials, such as polypropylene, Teflon®, cellulose and nitrocellulose membranes, microtiter plates (clear polystyrene plates; black, white, and clear polystyrene plates), glass slides and quartz optical fibers coated with 3-(amino propyl) triethoxy silane, etc., and several alternative protocols for achieving high monolayer coating densities. These studies have exposed phenomena neither evident nor of importance when antibody binding to solid supports is examined at a macroscopic level. Provisionally, we have used white Dynatech Microfluor microtiter plates—formulated for the detection of low fluorescence signals, and yielding high signal/noise ratios and high coating densities of functional antibodies (~5 × 10^4 IgG molecules/μm^2)—for assay development, although such plates are not ideal. Indeed, deficiencies in the antibody-deposition methods used constitute the principal source of imprecision in assay results and the limitation in sensitivity that this implies. Clearly, this represents an area for further study and refinement of current coating techniques.

Notwithstanding the limitations of present instrumentation (which, among other things, does not permit the use of time-resolving techniques to distinguish two individual fluorescence signals either from each other or from background fluorescence) and the crudeness of present methods for coupling antibodies onto small areas, we have verified the theoretical concepts outlined above by comparing the performance of several assays when constructed in microspot format and when conventionally designed. Although unoptimized, ratiometric microspot assays have yielded sensitivity values closely approaching those of conventional optimized IRMA. As an example, the results of a ratiometric assay system for thyrotropin, with use of Texas Red- and FITC-labeled antibodies, are shown in Figure 13. Bearing in mind the well-known limitations of these and other "conventional" fluorophors when used as immunoassay reagent labels, such results are encouraging, although further work is clearly required to achieve the considerably greater sensitivity theoretically predicted with use of improved fluorophors, better antibody-microspotting techniques, and purpose-built (time-resolving) instrumentation.

The finding that highly sensitive immunoassays can be performed with far smaller amounts of antibody than

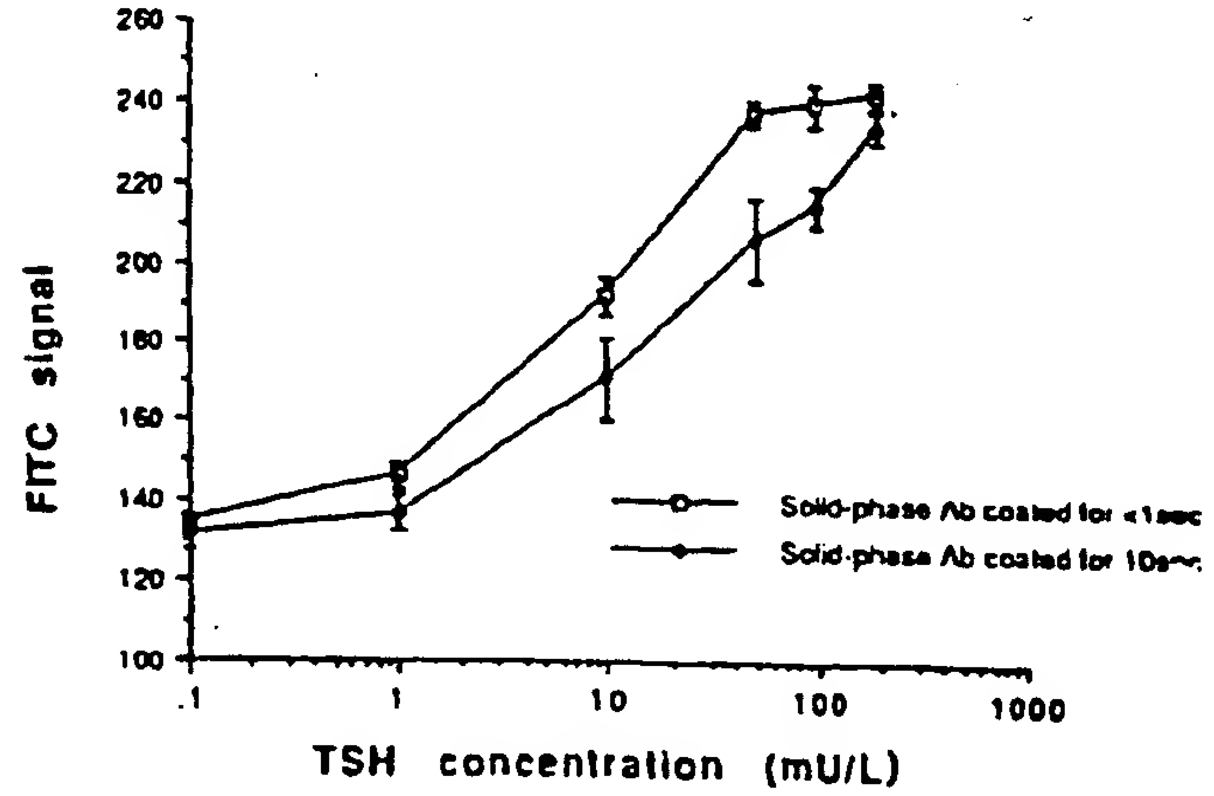


Fig. 13. Response curve in a dual-labeled microspot ratiometric assay of thyrotropin (TSH) with Texas Red-labeled solid-phase capture antibody and a developing antibody labeled with biotin/FITC–avidin

The FITC/Texas Red ratio for each microspot was measured with a scanning confocal microscope, and plotted as a function of TSH concentration in milli-int. units/L are currently used conventionally permits in turn the construction of antibody microspot arrays enabling, in principle, the simultaneous measurement of thousands of different substances in 1-mL samples. In collaboration with investigators at the Centre for Applied Microbiological Research, Porton Down, U.K., we are presently developing various techniques for the creation of such arrays. Indeed, similar technologies have recently been used for the parallel synthesis of several different polypeptides, these enabling 10 000-microspot arrays to be constructed on silica chips approximating 1 cm^2 (*24*). Although arrays of this capacity are unlikely to ever be required for conventional diagnostic purposes, we can anticipate that the ability to simultaneously measure many substances in the same sample will have revolutionary consequences in medicine and other similar areas. In addition, such techniques may ultimately permit the individual analysis of the multiple isoforms of certain "heterogeneous" analytes (e.g., the glycoprotein hormones), such molecular heterogeneity currently presenting a major obstacle to the standardization and interpretation of many immunological measurements (*25*). Moreover, although these concepts have been illustrated in an immunoassay context, they are clearly applicable to all "binding assays," including those relying on the use of DNA probes, hormone receptors, etc. For example, labeled lectins that are specific in their reactions with the sugar residues in the oligosaccharide chains of glycoprotein molecules may be used, together with specific antibodies, to impart additional "structural specificity" to sandwich assays (*26, 27*), possibly overcoming the limitations of antibodies per se in regard to differentiation of the glycosylation variants of the glycoprotein hormones.

Summary and Conclusion

Because of past confusion regarding the concepts of precision, sensitivity, accuracy, etc., several erroneous concepts have become incorporated within currently accepted rules of immunoassay design. In particular, much higher antibody concentrations are customarily used than are necessary to achieve very high assay sensitivity, provided that certain measurement strategies are adhered to. In this presentation, we have attempted to show that, in principle, the highest assay sensitivities are obtained by confining a small number of sensor antibody molecules onto a very small area in the form of a microspot and measuring their occupancy by an analyte, by using very high-specific-activity "developing" antibody probes, thereby maximizing the signal/noise ratio in the determination of sensor antibody occupancy. This observation, which contradicts currently accepted immunoassay design theory, in turn makes possible the measurement of an unlimited number of different analytes on a chip of very small surface area through the use of, e.g., laser scanning techniques closely analogous to those used in compact disk techniques of sound recording. Extensive experimental studies in this area, albeit conducted with relatively crude techniques and instrumentation not specifically designed for these purposes, and therefore not reported in detail here, have demonstrated the feasibility of the miniaturized antibody microspot approach and the validity of the general concepts on which it is based. We are therefore confident that this represents the basis of a next-generation technology that is likely to have a revolutionary impact on all fields involving the use of binding assays.

References

1. Yalow RS, Berson SA. General principles of radioimmunoassay. In: Hayes RL, Goswitz FA, Murphy BEP, eds. Radioisotopes in medicine: in vitro studies. Oak Ridge, TN: US Atomic Energy Commission, 1968:7–39.
2. Ekins RP, Newman B, O'Riordan JLH. *Ibid.*: 59–100.
3. Berson SA, Yalow RS. Measurement of hormones—radioimmunoassay. In: Berson SA, Yalow RS, eds. Methods in investigative and diagnostic endocrinology, Vol. 2A. Amsterdam: North Holland/Elsevier, 1973:84–135.
4. Ekins R, Newman B. Theoretical aspects of saturation analysis. In: Diczfalusy E, Diczfalusy A, eds. Steroid assay by protein binding. Karolinska symposia on research methods in reproductive endocrinology. Stockholm: WHO/Karolinska Sjukhuset, 1970:11–30.
5. Ekins RP. Limitations of specific activity. In: Margoulies M, ed. Protein and polypeptide hormones, Part 3 (Discussions). Amsterdam: Excerpta Medica, 1968:612–6, et seq.; Ekins RP. Concentrations of tracer and antiserum, time and temperature of incubation, volume of incubation. *Ibid:* 672–82.
6. Yalow RS, Berson SA. Immunoassay of endogenous plasma insulin in man. J Clin Invest 1960;39:1157.
7. Ekins RP. The estimation of thyroxine in human plasma by an electrophoretic technique. Clin Chim Acta 1960;5:453–9.
8. Barakat RM, Ekins RP. Assay of vitamin B_{12} in blood—a simple method. Lancet 1961;ii:25–6.
9. Wide L, Bennich H, Johansson SGO. Diagnosis of allergy by an in-vitro test for allergen antibodies. Lancet 1967;ii:1105–7.
10. Miles LEH, Hales CN. Labeled antibodies and immunological assay systems. Nature (London) 1968;219:186–9.
11. Keston AS, Udenfriend S, Cannan RK. Micro-analysis of mixtures (amino acids) in the form of isotopic derivatives. J Am Chem Soc 1946;68:1390.
12. Avivi P, Simpson SA, Tait JF, Whitehead JK. The use of ^{3}H and ^{14}C-labeled acetic anhydride as analytical reagents in microbiochemistry. In: Johnston JE, Faires RA, Millett RJ, eds. Radioisotope conference, London: Butterworths, 1954:313–23.
13. Miles LEH, Hales CN. An immunoradiometric assay of insulin. *Op. cit.* (ref. 5), Part 1:61–70.
14. Rodbard D, Weiss GH. Mathematical theory of immunometric (labeled antibody) assay. Anal Biochem 1973;52:10–44.
15. Jackson TM, Marshall NJ, Ekins RP. Optimisation of immunoradiometric assays. In: Hunter WM, Corrie JET, eds. Immunoassays for clinical chemistry. Edinburgh: Churchill Livingstone, 1983:557–75.
16. Ekins RP. Measurement of analyte concentration. British patent no. 8 224 600, 1983.
17. Wide L. Solid-phase antigen–antibody systems. In: Hunter WM, Kirkham KE, eds. Radioimmunoassay methods. Edinburgh: Churchill Livingstone, 1971:405–12.
18. Köhler G, Milstein C. Continuous culture of fused cells secreting specific antibody of predefined specificity. Nature (London) 1975;256:495–7.
19. Marshall NJ, Dakubu S, Jackson T, Ekins RP. Pulsed light, time resolved fluoroimmunoassay. In: Albertini A, Ekins RP, eds. Monoclonal antibodies and developments in immunoassay. Amsterdam: Elsevier/North Holland, 1981:101–8.
20. Soini E, Lövgren T. Time-resolved fluorescence of lanthanide probes and applications in biotechnology [Review]. Crit Rev Anal Chem 1987;18:105–54.
21. Ekins RP, Chu F, Biggart E. The development of microspot, multi-analyte ratiometric immunoassay using dual fluorescent-labeled antibodies. Anal Chim Acta 1990;227:73–96.
22. White JG, Amos WB, Fordham M. An evaluation of confocal versus conventional imaging of biological structures by fluores-

cence light microscopy. J Cell Biol 1987;105:41–8.
23. Ploem JS. New instrumentation for sensitive image analysis of fluorescence in cells and tissues. In: Tayer DL, Waggoner AS, Lanni F, Murphy R, Birge R, eds. Applications of fluorescence in the biological sciences. New York: Alan R Liss, 1986;289–300.
24. Fodor SPA, Read JL, Pirrung MC, et al. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–73.
25. Ekins RP. Immunoassay standardization. In: Kallner A, Magid E, Albert W, eds. Improvement of comparability and compatibility of laboratory assay results in life sciences. Immunoassay standardization. Scand J Clin Lab Invest 1991;51(Suppl 205):33–46.
26. Kottgen E, Hell B, Muller C, Tauber R. Demonstration of glycosylation variants of human fibrinogen, using the new technique of glycoprotein lectin immunosorbent assay (GLIA). Biol Chem Hoppe Seyler 1988;369:1157–66.
27. Kinoshita N, Suzuki S, Matsuda Y, Taniguchi N. α-Fetoprotein antibody–lectin enzyme immunoassay to characterise sugar chains for the study of liver diseases. Clin Chim Acta 1989;179:143–52.
28. Shalev V, Greenberg GH, McAlpine PJ. Detection of attograms of antigen by a high sensitivity enzyme-linked immunosorbent assay (HS-ELISA) using a fluorogenic substrate. J Immunol Methods 1980;38:125.
29. Harris CC, Yolken RH, Kroken H, Hsu IC. Ultrasensitive enzymatic radioimmunoassay: application to detection of cholera toxin and rotavirus. Proc Natl Acad Sci USA 1979;76:5336.

Corrections

Vol 37, pp. 1447–8: In our desire for rapid publication, important errors were introduced into the following Technical Brief. The corrected version is here reproduced in its entirety, with our apologies to the authors.

Rapid Detection of 1717-1G→A Mutation in CFTR Gene by PCR-Mediated Site-Directed Mutagenesis,

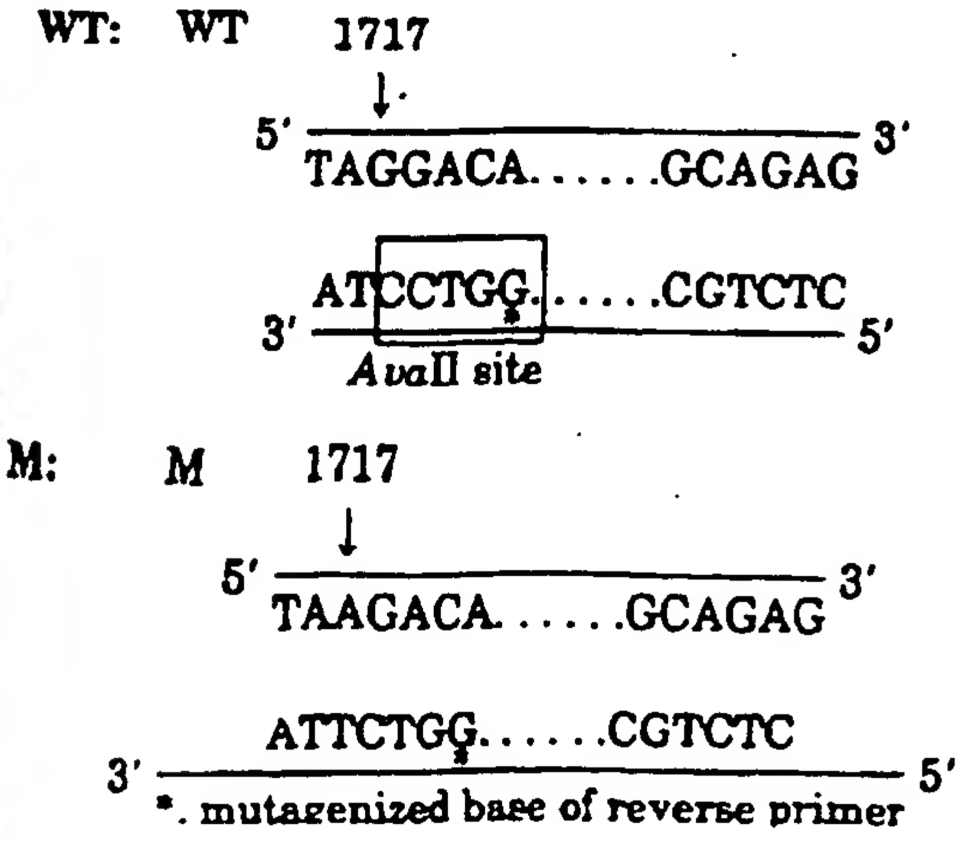
Laura Cremonesi,[1] *Manuela Seia,*[2] *Carmelina Magnani,*[1] *and Maurizio Ferrari*[1] ([1] Istituto Scientifico H.S. Raffaele, Lab. Centrale, Milano; [2] Istituti Clin. di Perfezionamento, Lab. di Ricerche Clin., Milano, Italy)

Until now, among the non-ΔF508 mutations identified in the cystic fibrosis transmembrane conductance regulator (CFTR) gene by the Cystic Fibrosis (CF) Genetic Analysis Consortium, the ones most frequently seen in our population sample are the 1717-1G→A mutation (13/144 or 9% of the CF chromosomes) and the G542X mutation (16/190 or 8.4% of the CF chromosomes), both revealed by dot–blot hybridization of the polymerase chain reaction (PCR) product with allele-specific oligonucleotides (ASO) probes (*1*).

In an attempt to simplify the analysis of the most frequent mutations in the CFTR gene, we converted radiolabeled ASO detection into restriction endonuclease analysis of the amplified product.

A PCR-mediated site-directed mutagenesis (*2, 3*) to detect the G542X mutation by generating a novel *BstNI* site in the wild-type sequence had already been suggested (*4*).

To detect the 1717-1G→A mutation, we designed the reverse primer (5′-CTCTGCAAACTTGGAGAGGTC-3′) to contain a single-base mismatch (T→G), which could create a novel *Ava*II restriction site [G ↓ G(A/T)CC] in the amplified wild-type (WT) allele but not in the CF mutant (M) allele:



*. mutagenized base of reverse primer

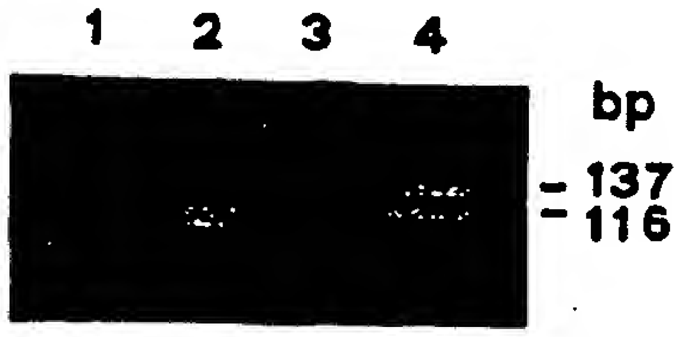


Fig. 1. Detection of the 1717-1G→A mutation by PCR

Reactions were carried out with 1 μg of genomic DNA in a total volume of 100 μL containing 10 mmol/L Tris · HCl (pH 8.3), 50 mmol/L KCl, 1.5 mmol/L $MgCl_2$, 0.1 g/L gelatin, 200 μmol/L each of the four deoxyribonucleotide triphosphates, 2.5 units of Taq polymerase (Perkin-Elmer Cetus, Norwalk, CT), and 100 pmol of each of the primers. PCR conditions were as follows: denaturation at 94 °C for 1 min, annealing at 55 °C for 30 s, and extension at 72 °C for 1 min, for a total of 30 cycles. PCR products were digested for 2 h at 37 °C with 5 U of AvaII and electrophoresed on 3% agarose–1% NuSieve gel for 1 h at 50 V. Bands were made visible by staining the gel with ethidium bromide. *Lane 1:* *Hae*III-digested pBR322 size marker. *Lane 2:* normal homozygote. *Lane 3:* CF patient homozygous for the 1717-1G→A mutation. *Lane 4:* heterozygote carrier for the 1717-1G→A mutation

For the forward primer, we used the one made available by the CF Genetic Analysis Consortium to amplify exon 11 of the CFTR gene: 5′-CAACTGTGGTTAAAGCAATAGTGT-3′.

Digestion by *Ava*II enzyme of the PCR product generates two fragments of 116- and 21-bp in the wild-type alleles and leaves undigested a 137-bp fragment in the mutant alleles (Figure 1).

By combined analysis for the ΔF508 mutation (*5*) (252/470 or 53.6% of the CF chromosomes), 1717-1G→A, and G542X, about 71% of mutations might be detected by nonisotopic analysis of the PCR product, thus allowing a faster and easier one-day procedure for carrier screening and prenatal testing.

References

1. Kerem B, Zielenski J, Markiewicz D, et al. Identification of mutations in regions corresponding to the two putative nucleotide (ATP)-binding folds of the cystic fibrosis gene. Proc Natl Acad Sci USA 1990;87:8447–51.
2. Haliassos A, Chomel JC, Baudis M, Kruh J, Kaplan JC, Kitzis A. Modification of enzymatically amplified DNA for the detection of point mutations. Nucleic Acids Res 1989;17:3606.
3. Friedman WE, Highsmith E Jr, Prior TW, Perry TR, Silverman LM. Cystic fibrosis deletion mutation detected by PCR-mediated site-directed mutagenesis [Tech Brief]. Clin Chem 1990;36:695–6.
4. Ng ISL, Pace R, Richard MV, et al. Methods for analysis of multiple cystic fibrosis mutations. Hum Genet (in press).
5. Ferrari M, Cremonesi L. More on detection of cystic fibrosis by polymerase chain reaction [Response to Letter]. Clin Chem 1990;36:1702–3.

clinical chemistry

11
91

In This Issue . . .

AMERICAN ASSOCIATION FOR CLINICAL CHEMISTRY
RESEARCH
SERVICE

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at http://www.pnas.org.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ≈0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

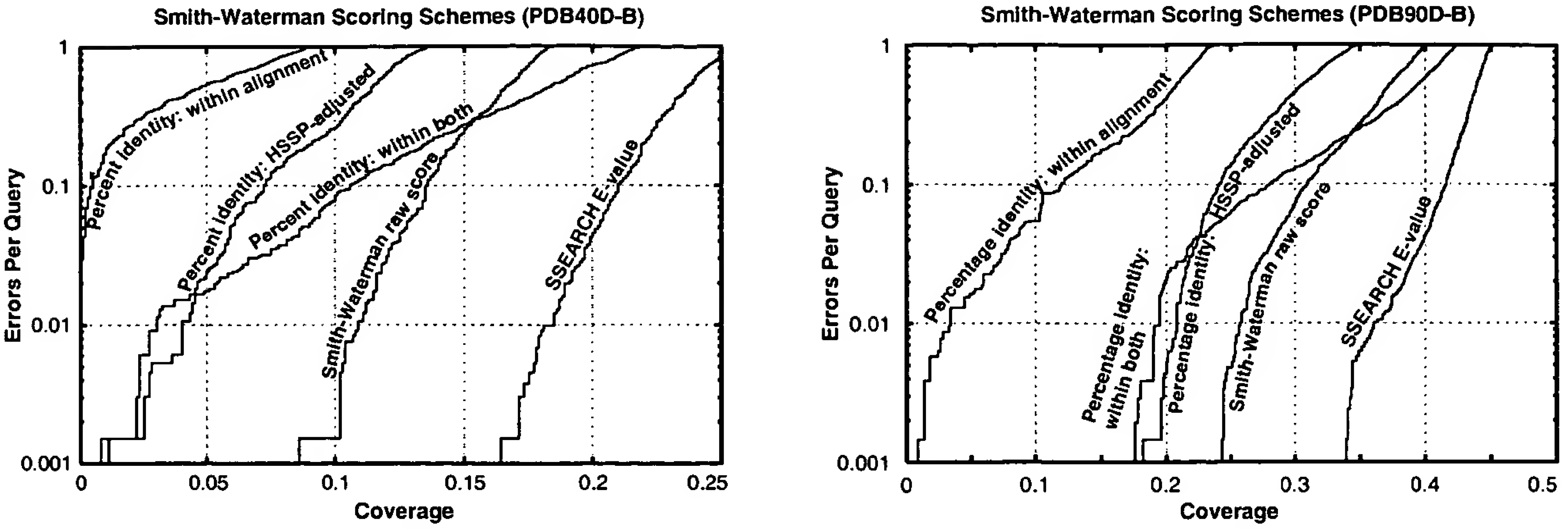
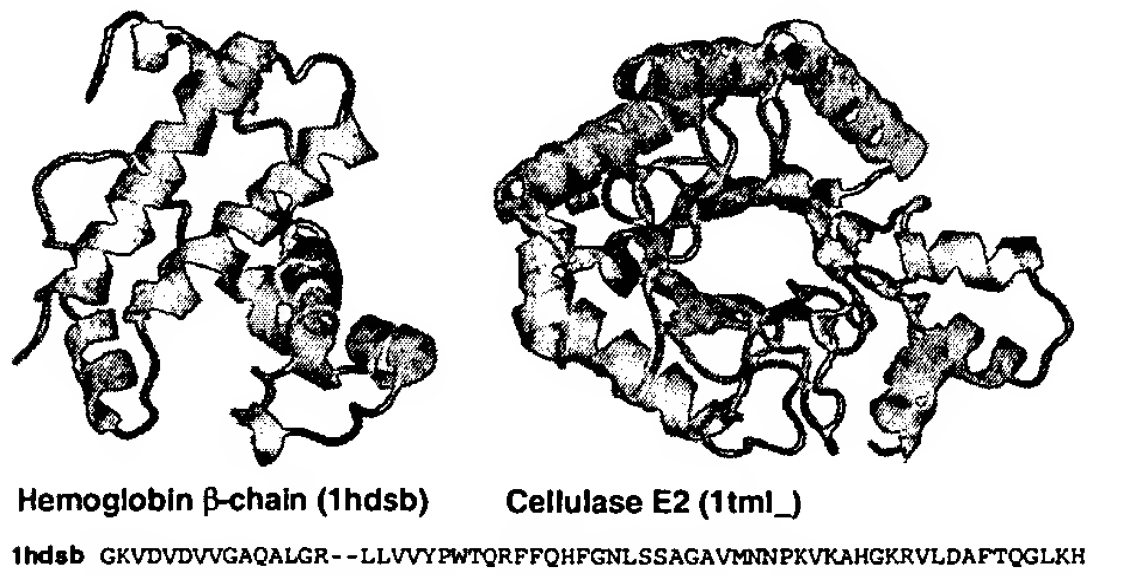


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the *x* axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The *y* axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The *y* axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely



```
1hdsb GKVDVDVVGAQALGR--LLVVYPWTQRFFQHFGNLSSAGAVMNNPKVKAHGKRVLDAFTQGLKH
      :.:.  .  .::: :.  .::::    :     :: ::.::    :. .:. .   .: :. :::.
1tml_ GQVDALMSAAQAAGKIPILVVYNAPGR---DCGNHSSGGA----PSHSAY-RSWIDEFAAGLKN
```

FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin *β*-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

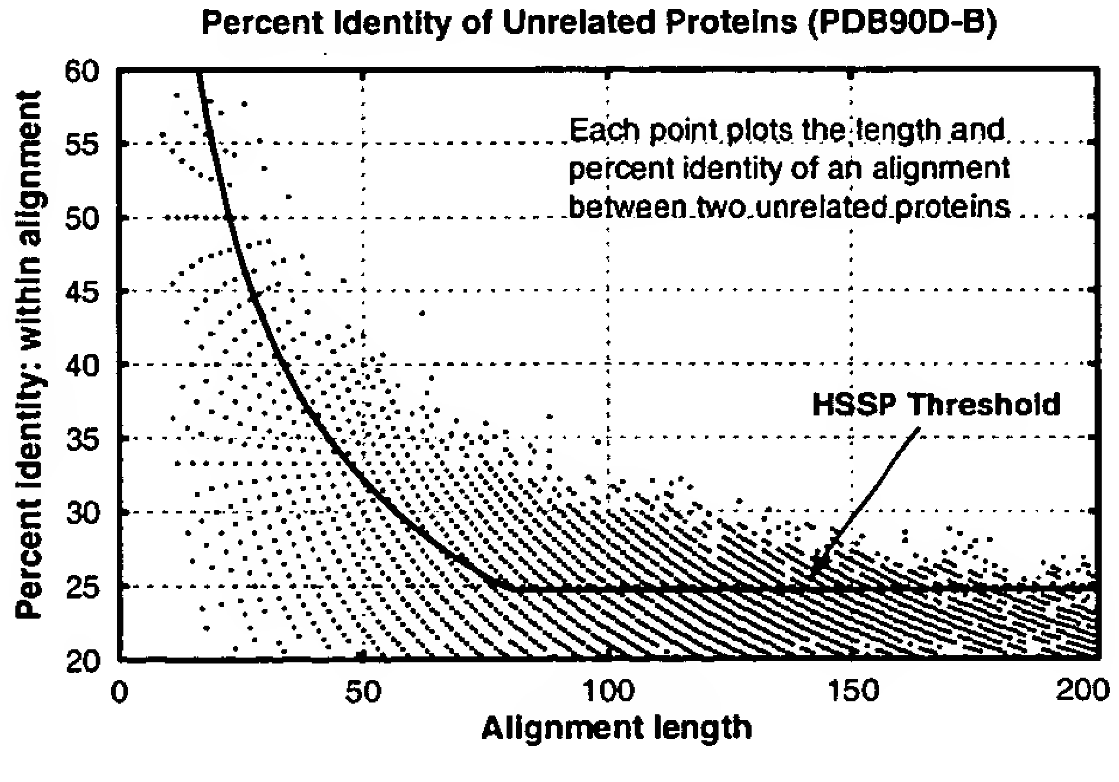


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

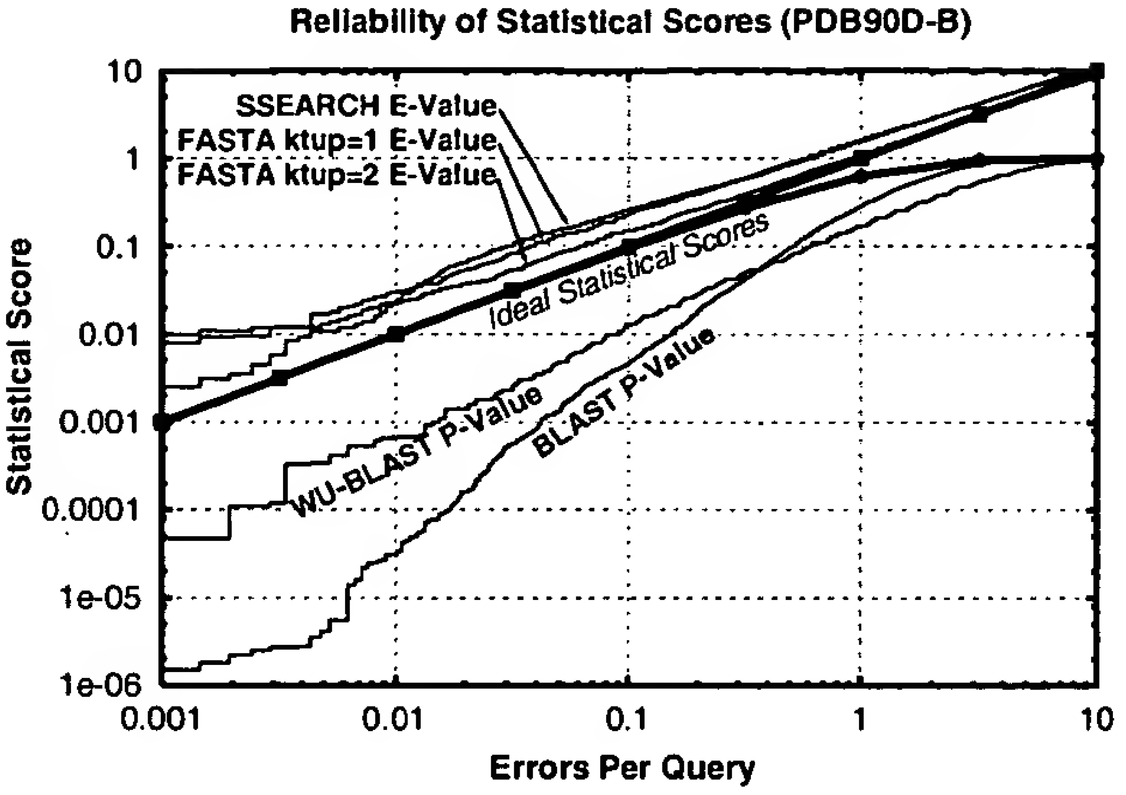


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

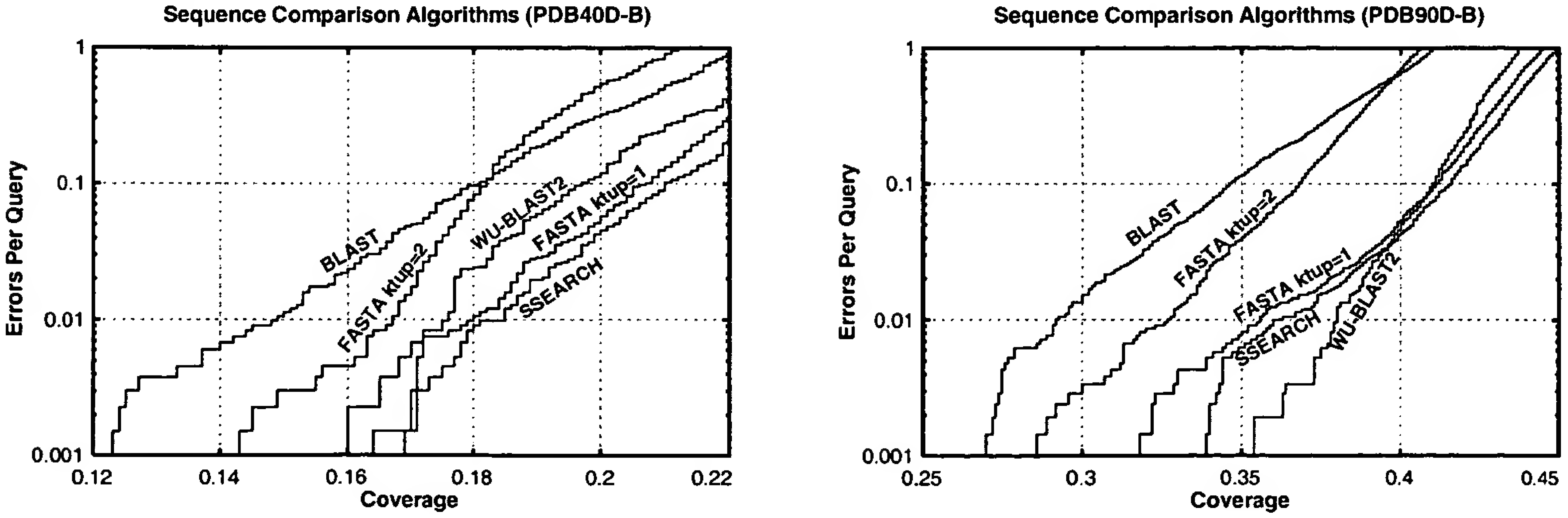


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

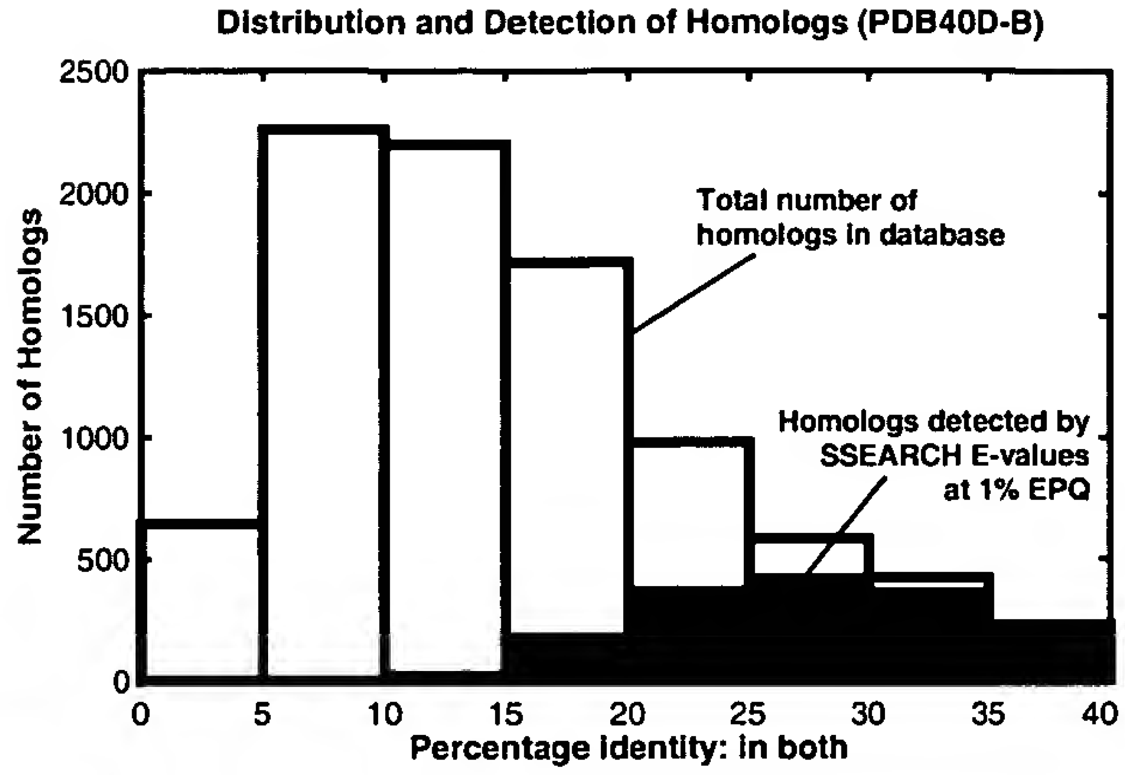


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266,** 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266,** 635–643.
6. Pearson, W. R. (1991) *Genomics* **11,** 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4,** 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266,** 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249,** 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17,** 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24,** 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24,** 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure,* ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233,** 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1,** 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1,** 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14,** 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6,** 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266,** 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12,** 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266,** 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9,** 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications,* eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7,** 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5,** 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39,** 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20,** 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9-16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4,** 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131,** 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32,** 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20,** 374–376.

JMB

The Relationship between Protein Structure and Function: a Comprehensive Survey with Application to the Yeast Genome

Hedi Hegyi and Mark Gerstein*

Department of Molecular Biophysics & Biochemistry Yale University, 266 Whitney Avenue, PO Box 208114 New Haven, CT, 06520 USA

*Corresponding author

For most proteins in the genome databases, function is predicted *via* sequence comparison. In spite of the popularity of this approach, the extent to which it can be reliably applied is unknown. We address this issue by systematically investigating the relationship between protein function and structure. We focus initially on enzymes functionally classified by the Enzyme Commission (EC) and relate these to by structurally classified domains the SCOP database. We find that the major SCOP fold classes have different propensities to carry out certain broad categories of functions. For instance, alpha/beta folds are disproportionately associated with enzymes, especially transferases and hydrolases, and all-alpha and small folds with non-enzymes, while alpha + beta folds have an equal tendency either way. These observations for the database overall are largely true for specific genomes. We focus, in particular, on yeast, analyzing it with many classifications in addition to SCOP and EC (i.e. COGs, CATH, MIPS), and find clear tendencies for fold-function association, across a broad spectrum of functions. Analysis with the COGs scheme also suggests that the functions of the most ancient proteins are more evenly distributed among different structural classes than those of more modern ones. For the database overall, we identify the most versatile functions, i.e. those that are associated with the most folds, and the most versatile folds, associated with the most functions. The two most versatile enzymatic functions (hydro-lyases and *O*-glycosyl glucosidases) are associated with seven folds each. The five most versatile folds (TIM-barrel, Rossmann, ferredoxin, alpha-beta hydrolase, and P-loop NTP hydrolase) are all mixed alpha-beta structures. They stand out as generic scaffolds, accommodating from six to as many as 16 functions (for the exceptional TIM-barrel). At the conclusion of our analysis we are able to construct a graph giving the chance that a functional annotation can be reliably transferred at different degrees of sequence and structural similarity. Supplemental information is available from http://bioinfo.mbb.yale.edu/genome/foldfunc.

Keywords: structure-function; fold classification; structural convergence; functional divergence; yeast genomics

Introduction

The problem of determining function from sequence

An ultimate goal of genome analysis is to determine the biological function of all the gene products in a genome. However, the function of only a minor fraction of proteins has been studied experimentally, and, typically, prediction of function is based on sequence similarity with proteins of known function. That is, functional annotation is transferred based on similarity. Unfortunately, the relationship between sequence similarity and functional similarity is not as straightforward. This has been commented on in numerous reviews (Bork & Koonin, 1998; Karp, 1998). Karp (1998), in particular, has noted that transferring of incorrect functional information threatens to

Abbreviations used: EC, Enzyme Commission; ORF open reading frame.

E-mail address of the corresponding author: Mark.Gerstein@yale.edu

0022-2836/99/160147-18 $30.00/0

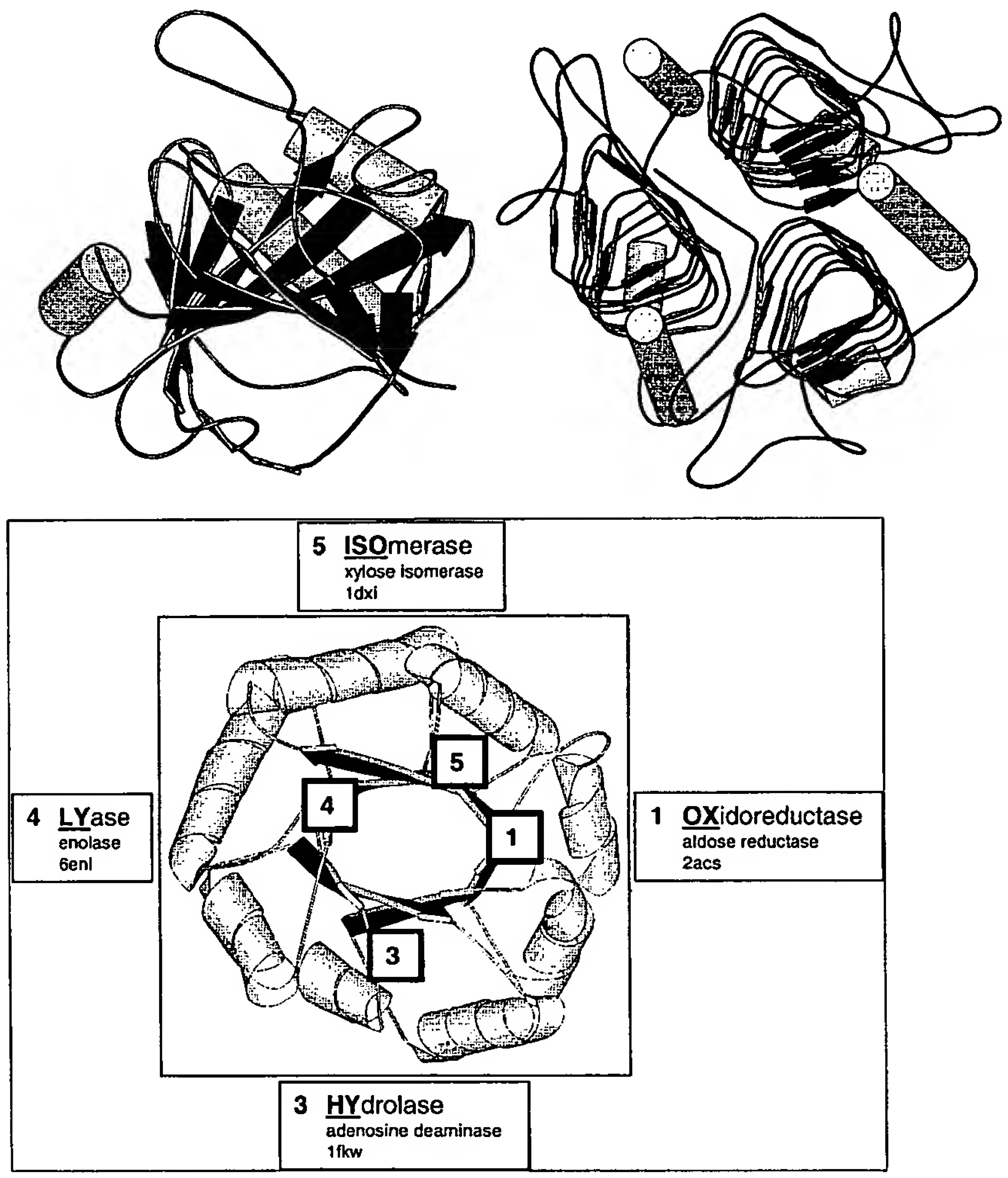


Figure 1. Specific example of convergent and divergent evolution. Top, an example of convergent evolution, showing structures of two carbonic anhydrases with the same enzymatic function (EC number 4.2.1.1), but with different folds. The Figure was drawn with Molscript (Kraulis, 1991) from 1THJ (left-handed beta helix) and 1DMX (flat beta sheet). Bottom, an example of possible divergent evolution, the TIM-barrel. This fold functions as a generic scaffold catalyzing 15 different enzymatic functions. A schematic Figure of the TIM-barrel fold is shown with numbers in boxes indicating the different location of the active site in four proteins that have this fold. These four proteins, xylose isomerase, aldose reductase, enolase, and adenosine deaminase, carry out very different enzymatic functions, in four of the main EC classes (1.*.*, 3.*.*, 4.*.*, and 5.*.*). They have active sites at very different locations (identified by the boxed numbers in the barrel) yet they all share the same fold.

progressively corrupt genome databases through the problem of accumulating incorrect annotations and using them as a basis for further annotations, and so on.

It is known that sequence similarity does confer structural similarity. Moreover, there is a well-established quantified relationship between the extent of similarity in sequence and that in structure. First investigated by Chothia & Lesk (1986) the similarity between the structures of two proteins (in terms of RMS) appears to be a monotonic function of their sequence similarity. This fact is often exploited when two sequences are declared related, based on a database search by programs such as BLAST or FastA (Altschul *et al.*, 1997; Pearson, 1996). Often, the only common element in two distantly related protein sequences is their underlying structures, or folds.

Transitivity requires that the well-established relationship between sequence and structure, and the more indefinite one between sequence and function, imply an indefinite relationship between structure and function. Several recent papers have highlighted this, analyzing individual protein superfamilies with a single fold but diverse functions. Examples include the aldo-keto reductases, a large hydrolase superfamily, and the thiol protein esterases. The latter include the eye-lens and corneal crystallins, a remarkable example of functional divergence (Bork & Eisenberg, 1998; Bork *et al.*, 1994; Cooper *et al.*, 1993; Koonin & Tatusov, 1994; Seery *et al.*, 1998).

There are also many classic examples of the converse: the same function achieved by proteins with completely different folds. For instance, even though mammalian chymotrypsin and bacterial subtilisin have different folds, they both function as serine proteases and have the same Ser-Asp-His catalytic triad. Other examples include sugar kinases, anti-freeze glycoproteins, and lysyl-tRNA synthetases (Bork *et al.*, 1993; Chen *et al.*, 1997; Doolittle, 1994; Ibba *et al.*, 1997a,b).

Figure 1 shows well-known examples of each of these two basic situations: the same fold but different function (divergent evolution) and the same function but different fold (convergent evolution).

Protein classification systems

The rapid growth in the number of protein sequences and three-dimensional structures has made it practical and advantageous to classify proteins into families and more elaborate hierarchical systems. Proteins are grouped together on the basis of structural similarities in the FSSP (Holm & Sander, 1998), CATH (Orengo *et al.*, 1997), and SCOP databases (Murzin *et al.*, 1995). SCOP is based on the judgments of a human expert FSSP, on automatic methods, and CATH, on a mixture of both. Other databases collect proteins on the basis of sequence similarities to one another, e.g. PROSITE, SBASE, Pfam, BLOCKS, PRINTS and ProDom (Attwood *et al.*, 1998; Bairoch *et al.*, 1997; Corpet *et al.*, 1998; Fabian *et al.*, 1997; Henikoff *et al.*, 1998; Sonnhammer *et al.*, 1997). Several collections contain information about proteins from a functional point of view. Some of these focus on particular organisms, e.g. the MIPS functional catalogue and YPD for yeast (Mewes *et al.*, 1997; Hodges *et al.*, 1998) and EcoCyc and GenProtEC for *Escherichia coli* (Karp *et al.*, 1998; Riley, 1997). Others focus on particular functional aspects in multiple organisms, e.g. the WIT and KEGG databases, which focus on metabolism and pathways (Selkov *et al.*, 1997; Ogata *et al.*, 1999), the ENZYME database, which focuses obviously enough on enzymes (Bairoch, 1996), and the COGs system, which focuses on proteins conserved over phylogenetically distinct species (Tatusov *et al.*, 1997). The ENZYME database, in particular, contains all the enzyme reactions that have an Enzyme Commission (EC) number assigned in accordance with the International Nomenclature Committee and is cross-referenced with Swissprot (Bairoch, 1996; Bairoch & Apweiler, 1998; Barrett, 1997).

Our approach: systematic comparison of proteins classified by structure with those classified by function

One of the most valuable operations one can do to these individual classification systems is to cross-reference and cross-tabulate them, seeing how they overlap. We performed such an analysis here by systematically interrelating the SCOP, Swissprot and ENZYME databases (Bairoch, 1996; Bairoch & Apweiler, 1998; Murzin *et al.*, 1995). For yeast we also have used the MIPS yeast functional catalogue, CATH and COGs in our analysis. This enables us to investigate the relationship between protein function and structure in a comprehensive statistical fashion. In particular, we investigated the functional aspects of both divergent and convergent evolution, exploring cases where a structure gains a dramatically different biochemical function and finding instances of similar enzymatic functions performed by unrelated structures.

We concentrated on single-domain Swissprot proteins with significant sequence similarity to one of the SCOP structural domains. Since most of these proteins have a single assigned function, comparing them to individual structural domains, which can have only one assigned fold, allowed us to establish a one-to-one relationship between structure and function.

Recent related work

This work is following up on several recent reports on the relationship between protein structure and function. In particular, Martin *et al.* (1998) studied the relationship between enzyme function and the CATH fold classification. They concluded that functional class (expressed by top-level EC numbers) is not related to fold, since a few specific residues, not the whole fold, determine enzyme function. Russell (1998) also focused on specific side-chain patterns, arguing that these could be used to predict protein function. In a similar fashion, Russell *et al.* (1998) identified structurally similar "supersites" in superfolds. They estimated that the proportion of homologues with different binding sites, and therefore with different functions, is around 10%. In a novel approach, using machine learning techniques, des Jardins *et al.* (1997) predict purely from the sequence whether a given protein is an enzyme and also the enzyme class to which it belongs.

Our work is also motivated by recent work looking at whether or not organisms are characterized by unique protein folds (Frishman & Mewes, 1997; Gerstein, 1997, 1998a,b; Gerstein & Hegyi, 1998; Gerstein & Levitt, 1997). If function is closely associated with fold (in a one-to-one sense), one would think that when a new function arose in evolution, nature would have to invent a new fold. Conversely, if fold and function are only weakly coupled, one would expect to see a more uniform distribution of folds amongst organisms and a high incidence of convergent evolution. In fact, a recent study on microbial genome analysis claims that functional convergence is quite common (Koonin & Galperin, 1997). Another related paper systematically searched Swissprot for all such cases of what is termed "analogous" enzymes (Galperin *et al.*, 1998).

Our work is also motivated by the recent work on protein design and engineering which aims to rationally change a protein function, for instance, to engineer a reporter function into a binding protein (Hellinga, 1997, 1998; Marvin *et al.*, 1997).

Results

Overview of the 8937 single-domain matches

Our basic results were based on simple sequence comparisons between Swissprot and SCOP, the SCOP domain sequences being used as queries against Swissprot. We focused on "mono-functional" single-domain matches in Swissprot, i.e. those singe-domain proteins with only one annotated function. The detailed criteria used in the database searches are summarized in Materials and Methods.

Overall, a little more than a quarter of the proteins in Swissprot are enzymes, a similar fraction are of known structure, and about one-eighth are both. (More precisely, of the 69,113 analyzed proteins in Swissprot, 19,995 are enzymes, 18,317 are structural homologues, and 8205 are both.) About half of the fraction of Swissprot that matched known structures were "single-domain" and about one-third of these were enzymes (8937 and 3359, respectively, of 18,317). We focus on these 8937 single-domain matches here. Notice how these numbers also show how the known structures are significantly biased towards enzymes: 45% (8205 out of 18,317) of all the structural homologues are enzymes *versus* 29% (19,995 out of 69,113) for all of Swissprot.

331 observed fold-function combinations

Figure 2 gives an overview of how the matches are distributed amongst specific functions and folds. The single-domain matches include 229 of the 361 folds in SCOP 1.35, and 91 of the 207 three-component enzyme categories in the ENZYME database (Bairoch, 1996). Each match combines a SCOP fold number on the structural side (columns in Figure 2) and a three-component EC category on the functional side (rows), with all the non-enzymatic functions grouped together into a single category with the artificial "EC number" of 0.0.0 (shown in the first row in Figure 2). This results in a table where each cell represents a potential fold-function combination. The table contains a maximum of 21,068 (=229 x 92) possible fold-function combinations (and a minimum of 229 combinations, assuming only one function for every fold). We actually observe 331 of these combinations (1.6%, shown by the filled-in cells).

Overall, more than half of the functions are associated with at least two different folds, while less than half of the folds with enzymatic activity have at least two functions (51 out of 91 and 53 out of 128, respectively).

Summarizing the fold-function combinations by 42 broad structure-function classes

As listed in Table 1, folds can be subdivided in six broad fold classes (e.g. all-alpha, all-beta, alpha/beta, etc.). Likewise, functions can be broken into seven main classes, non-enzymes plus six enzyme classes, e.g. oxidoreductase, transferase, etc. This gives rise to 42 (6 x 7) structure-function classes. The way the 21,068 potential fold-function combinations are apportioned amongst the 42 classes is shown in Table 2A.

Table 2B shows the way the 331 observed combinations were actually distributed amongst the 42 classes. Comparing the number of possible combinations with that observed shows that the most densely populated region of the chart is the transferase, hydrolase and lyase functions in combi-

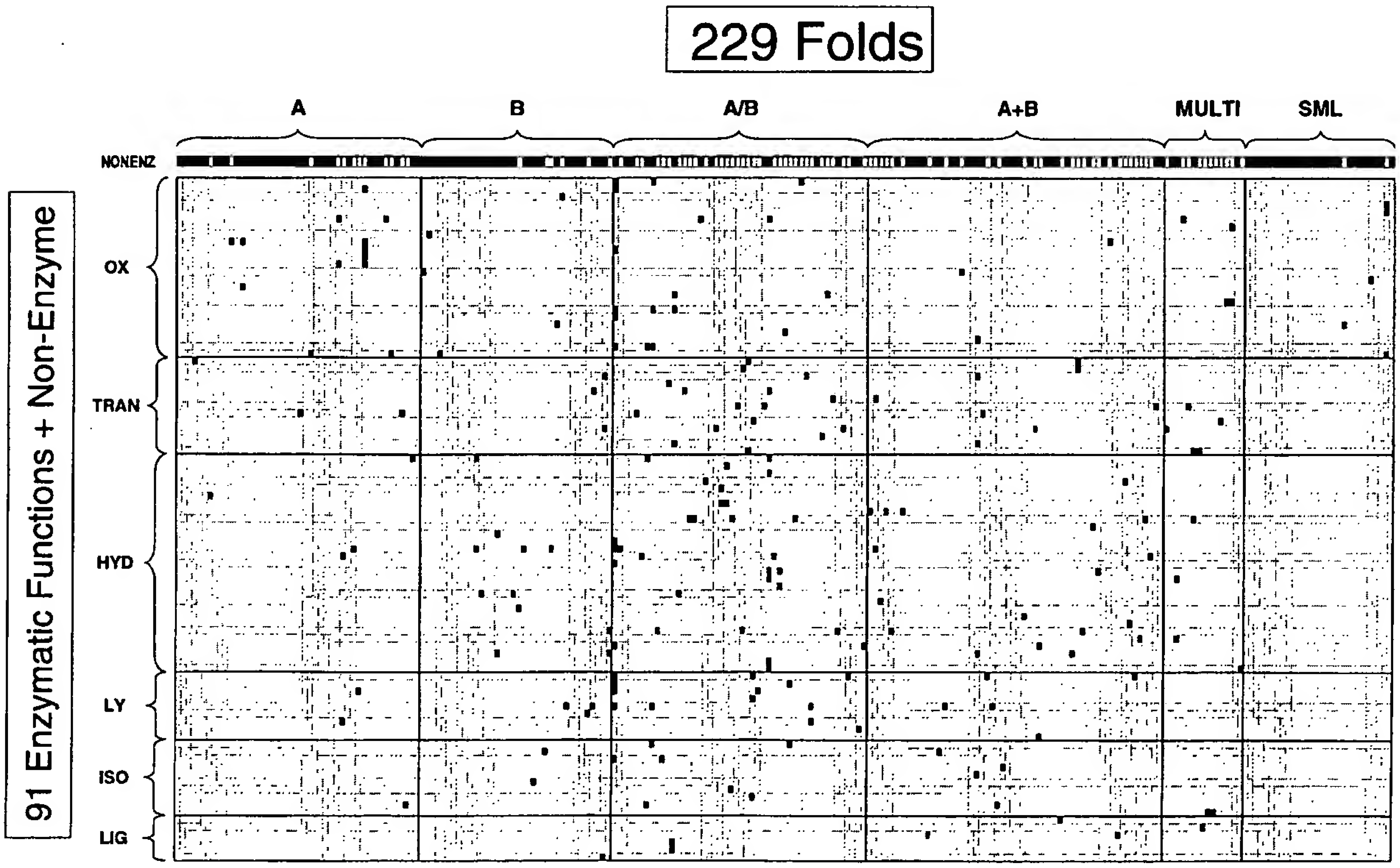


Figure 2. Overview of all the single-domain matches between proteins in Swissprot 35 and domains in SCOP 1.35. Sequences were compared with BLAST using the match criteria described in Materials and Methods. The matches are clustered into 92 functions (based on three-component EC numbers), which are arranged on each row, and 229 folds (based on SCOP fold numbers), which are arranged on each column. The first row indicates the matches with non-enzymes. There are, thus, 21,068 (=92 x 229) possible combinations shown in the Figure. Only the 331 are actually observed. These are indicated by filled squares.

Table 1. Broad structural and functional categories

A. *Functional categories in Swissprot 35*[a]

EC category	Category name	Abbreviation	Num. of functions in category
0.0.0	Non-enzymes	NONENZ	1
1.*.*	Oxidoreductases	OX	86
2.*.*	Transferases	TRAN	28
3.*.*	Hydrolases	HYD	53
4.*.*	Lyases	LY	15
5.*.*	Isomerases	ISO	16
6.*.*	Ligases	LIG	9
		Total:	208

B. *Structural classes in SCOP 1.35*[b]

Fold class	Class name	Abbreviation	Num. of folds in class
1	All-alpha	A	81
2	All-beta	B	57
3	Alpha and beta	A/B	70
4	Alpha plus beta	A + B	91
5	Multi-domain	MULTI	19
6	Transmembrane	TM	9
7	Small proteins	SML	43
		Total:	361

[a] List of the functional (enzymatic) categories in Swissprot and the abbreviations used here. The values denote the number of three-component EC numbers in each category.

[b] List of the structural classes in SCOP studied here, and the abbreviations used for the classes. Values denote the number of folds in each class in SCOP 1.35. Class 6 is not used in the analysis.

nation with the alpha/beta fold class. This notion is in accordance with the general view that the most popular structures among enzymes fall into the alpha/beta class. In contrast, matches between small folds and enzymes are almost completely missing, except for five folds in the oxidoreductase category. There are also no all-alpha ligases and only one all-alpha isomerase.

Table 2C and D break down the 331 fold-function combinations in Table 2A into either just a number of folds or just a number of functions. That is, Table 2C lists the number of different folds associated with each of the 42 structure-function classes (corresponding to the non-zero columns in the relevant class in Figure 2), and Table 2D does the same thing for functions (non-zero rows in Figure 2). Comparing these tables back to the total number of combinations (Table 2A) reveals some interesting findings, keeping in mind that more functions than folds reveals probable divergence and that more folds than functions reveals probable convergence. For instance, the alpha/beta and alpha + beta fold classes contain similar numbers of folds, but the alpha/beta class has relatively more functions, perhaps reflecting a greater divergence. (Specifically, the alpha/beta class has 73 folds and 56 functions, while the alpha + beta class has 67 folds but only 35 functions.)

Table 2E shows the number of matching Swissprot sequences (from the total of 69,113) for each of the 42 structure-function classes. The most highly populated categories are the all-alpha non-enzymes, where 683 of the 1940 matches come from globins, and the all-beta non-enzymes, where 361 of the 1159 Swissprot sequences have matches with the immunoglobulin fold. These numbers are, obviously, affected by the biases in Swissprot. On the other hand, if we compare the total matches in Table 2E with the total combinations in Table 2B it is clear that the numbers do not directly correlate. For instance, fewer hydrolases in Swissprot have matches with alpha/beta folds than with alpha + beta folds (295 *versus* 452), but the number of different combinations in the first case is 30, as opposed to only 18 in the second case. This suggests that our approach of counting combinations may not be as affected by the biases in the databanks as simply counting matches.

Table 2F and G give some rough indication of the statistical significance of the differences in the observed distribution of combinations. In Table 2F, using chi-squared statistics, we calculate for each individual structure class the chance that we could get the observed distribution of fold-function combinations over various functional classes if fold was not related to function. Then in Table 2G, we reverse the role of fold and function, and calculate the statistics for each functional class.

Enzyme *versus* non-enzyme folds

On the coarsest level, function can be divided amongst enzymes and non-enzymes. Of the 229 folds present in Figure 2, 93 are associated only with enzymes and 101 are associated only with non-enzymes. The remaining folds were associated with both enzymatic and non-enzymatic activity. Finally, of the 93 purely enzymatic folds, 18 have multiple enzymatic functions.

Figure 3(a) shows a graphical view of the distribution of the different fold classes among these broadest functional categories. The distribution is far from uniform. The all-alpha fold class has 30 non-enzymatic representatives, but only 12 purely enzymatic folds and four folds with "mixed" (both types of) functions. This implies that a protein with an all-alpha fold has *a priori* roughly twice the chance of having a non-enzymatic function over an enzymatic one. The all-beta fold class has six enzymatic, 17 non-enzymatic and 13 mixed folds. In the alpha/beta class, 34 folds are associated only with enzymes and five folds only with non-enzymes, whereas in the alpha + beta class this ratio is more balanced, 28 "purely" enzymatic folds *versus* 22 purely non-enzymatic ones.

Table 2. Statistics over 42 structure-function classes

A. *Number of possible combinations between folds and functions in each of 42 classes (number of cells in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	46	36	48	56	15	28	229
OX	1104	864	1152	1344	360	672	5496
TRAN	598	468	624	728	195	364	2977
HYD	1334	1044	1392	1624	435	812	6641
LY	414	324	432	504	135	252	2061
ISO	460	360	480	560	150	280	2290
LIG	276	216	288	336	90	168	1374
Sum	4232	3312	4416	5152	1380	2576	21,068

B. *Number of observed combinations between folds and functions in each of 42 classes (number of filled cells in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	34	30	14	28	4	26	136
OX	13	5	17	3	4	5	47
TRAN	3	3	16	8	5		35
HYD	4	11	30	18	4		67
LY	2	3	13	5			23
ISO	1	2	7	4	2		16
LIG		1	2	3	1		7
Sum	57	55	99	69	20	31	331

C. *Number of folds in each of the 42 classes (columns with a filled cell in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	34	30	14	28	4	26	136
OX	7	5	9	3	3	3	30
TRAN	3	2	15	6	5		31
HYD	4	8	19	18	3		52
LY	2	3	8	5			18
ISO	1	2	7	4	2		16
LIG		1	1	3	1		6
Sum	51	51	73	67	18	29	289

D. *Number of functions in each of the 42 classes (rows with a filled cell in Figure 2)*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	1	1	1	1	1	1	6
OX	8	5	9	3	3	5	33
TRAN	2	3	13	8	4		30
HYD	4	7	19	14	4		48
LY	2	2	7	3			14
ISO	1	2	5	4	1		13
LIG		1	2	2	1		6
Sum	18	21	56	35	14	6	150

E. *Total number of matching Swissprot sequences in each of the 42 fold-function classes*

	A	B	A/B	A + B	MULTI	SML	Sum
NONENZ	1940	1159	560	638	106	892	5295
OX	150	202	388	50	68	18	876
TRAN	65	14	363	116	174		732
HYD	116	394	295	452	92		1349
LY	40	47	168	104			359
ISO	2	54	122	22	2		202
LIG		5	26	69	24		124
Sum	2313	1875	1922	1451	466	910	8937

F. *How much does each of the fold classes deviate from the average distribution of functions?*

	χ^2	P
A	17.5	<0.01
B	5.2	<0.6
A/B	32.5	<0.00002
A + B	7.7	<0.3
MULTI	9.9	<0.2
SML	27.8	<0.0002

continued

Table 2—*Continued*

G. *How much do each of the function classes deviate from the average distribution of folds?*

	χ^2	P
NONENZ	40.7	<0.0000002
OX	9.9	<0.08
TRAN	13.1	<0.03
HYD	17.3	<0.005
LY	10.2	<0.08
ISO	5.0	<0.5
LIG	4.3	<0.6

This Table shows various totals from Figure 2 distributed among the 42 structure-function classes, i.e. the seven functional categories in Table 1A multiplied by the six structural categories in Table 1B. Part A shows how many potential fold-function combinations there are in Figure 2 amongst each of the 42 classes. Part B shows how many of these 21,068 possible combinations are actually observed. Part C shows the total number of different folds (i.e. selected columns in Figure 1) in each class. Part D shows the total number of different functions (i.e. selected rows in Figure 2) in each class. Part E shows the total number of matching Swissprot proteins in the 42 classes. Note that to observe a fold-function combination one only needs the existence of a single match between a Swissprot protein and a SCOP domain. However, there can be many more. That is why the totals in this Table sum up to so much larger an amount than 331.

Here is an example of how to read parts A to E of the Table, focussing on the all-alpha, oxidoreductase region. Part A shows that there are 1104 cells, filled or unfilled, in this region, corresponding to possible combinations. Part B shows that 13 of these 1104 cells are filled, corresponding to observed all-alpha, oxidoreductase combinations. Part C shows that there are seven folds, corresponding to columns with filled cells in this region. Part D shows that there are eight functions, corresponding to rows with filled cells in this region. Finally, in part E we find that there are 150 Swissprot entries that have matches with a SCOP domain. They correspond to the 13 observed combinations in Part B.

Parts F and G give information on the statistical significance of the differences observed between the 42 structure-function classes. Part F gives the significance that the observed distribution of fold-function combinations in a given functional class is different than average (i.e. the null hypothesis that distribution of fold-function combinations is the same in each functional class). This is very similar to the derivation by Martin *et al.* (1998). A chi-squared statistic is computed for each of the seven functional classes in the conventional way: $\chi^2(f) = \Sigma_s (O_{sf} - E_{sf})^2/E_{sf}$, where for a given functional class f and structure class s, O_{sf} is the observed number of fold-function combinations and E_{sf} is the expected number. E_{sf} is simply computed from scaling the "sum" column and row in Part B of the Table: $E_{sf} = T_s T_f/T$, where T_s is the total number of combinations in a given structural class s (sum row), T_f is the total number of combinations in a given functional class f (sum column), and T is the total observed number of combinations, 331. Part G gives the statistical significance that the observed distribution of fold-function combinations in a given structural class is different than average. To compute this one simply sums over functions instead of structures: $\chi^2(s) = \Sigma_f(O_{sf} - E_{sf})^2/E_{sf}$. After each chi-squared statistic is reported, a rough probability or P-value is given. This gives the chance the observed distribution could be obtained randomly.

Restricting the comparison to individual genomes

Figure 3(a) applies to all of Swissprot. Figure 3(b) and (c) shows the functional distribution of folds taking into account the matches only in two specific genomes, yeast and *E. coli*. Only a fraction of each genome could be taken into consideration for various reasons (156 proteins in yeast, 244 proteins in *E. coli*), mostly due to the great number of enzymes having multiple domains in both yeast and *E. coli*. Chi-squared tests show that the fold distribution in yeast does not differ significantly from that in Swissprot and that the one in *E. coli* differs only slightly ($P < 0.25$ and $P < 0.02$, respectively). The main difference between Swissprot and *E. coli* is the larger fraction of alpha/beta enzymatic folds in the latter (34/93 *versus* 26/49). There are also somewhat more non-enzymatic all-alpha and small folds in Swissprot than in the two genomes. This is principally due to the greater prevalence of globins, myosins, cytochromes, toxins, and hormones in Swissprot than in yeast and *E. coli*. Many of these, of course, are proteins usually associated with multicellular organisms. We did a preliminary version of the fold distribution for the worm *Caenorhaditis elegans*. As expected this distribution turns out to be similar to that of Swissprot (data not shown).

The yeast genome viewed from different classification schemes

In Figure 4 we focus on the yeast genome in more detail, trying to see the effect that different classification schemes have on our results. Although the total number of counts for our statistics decrease, in just using yeast relative to all of Swissprot, yeast provides a good reference frame to compare a number of classification schemes in as unbiased a fashion as possible. Also, yeast is one of the most comprehensively characterized organisms, and there are a number of functional classifications available exclusively for this organism.

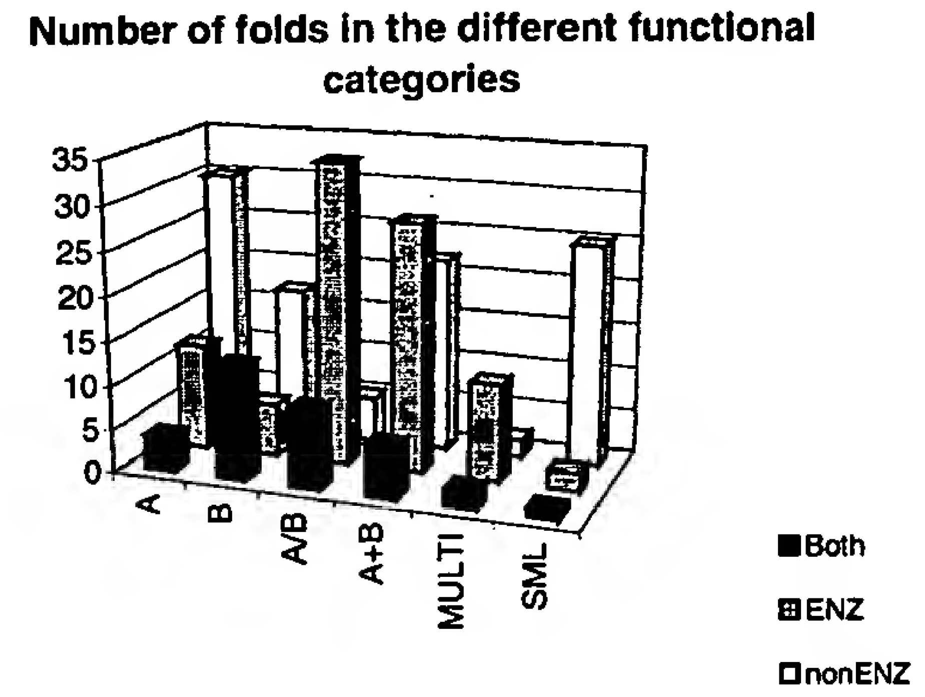
In part Figure 4(a) we cross-tabulate the structure-function combinations in yeast using the SCOP and EC systems as we have done for all of Swissprot in Table 2B. The yeast distribution is fairly similar to that of Swissprot with the only major difference being somewhat more alpha/beta transferases and fewer alpha/beta hydrolases than expected. (A chi-squared test gives $P < \sim 0.05$ for the two distributions to differ. If either the transferase or hydrolase difference is removed, P increases to ~20%.)

Figure 4(b) shows the structure-function combinations based on using the CATH structural classification (Orengo *et al.*, 1997) instead of SCOP. For this Figure we mapped the SCOP classification of a yeast PDB match to its corresponding CATH classification and then cross-tabulated the structure-function combinations in the various classes. Essentially, this Figure shows the results reported by Martin *et al.* (1998) just for yeast.

In Figure 4(c) and (d), which show COGs *versus* SCOP cross-tabulations, we achieve the opposite of (b). We change the functional classifications scheme but keep SCOP for classifying structures. As was the case with the enzyme classification, but perhaps even more so, using COGs to classify function shows clearly that certain fold classes are associated with certain functions and *vice versa*. Most notably, whereas the functions associated with metabolism, which are mostly enzymes, are preferentially associated with the alpha/beta fold class, those associated with cellular processes (e.g. secretion) and information processing (e.g. transcription), show no such preference. They, in fact, show a marked preference for all-alpha structure. Small proteins are absent from most of the COGs classes, except one part of information processing and two in cellular processes.

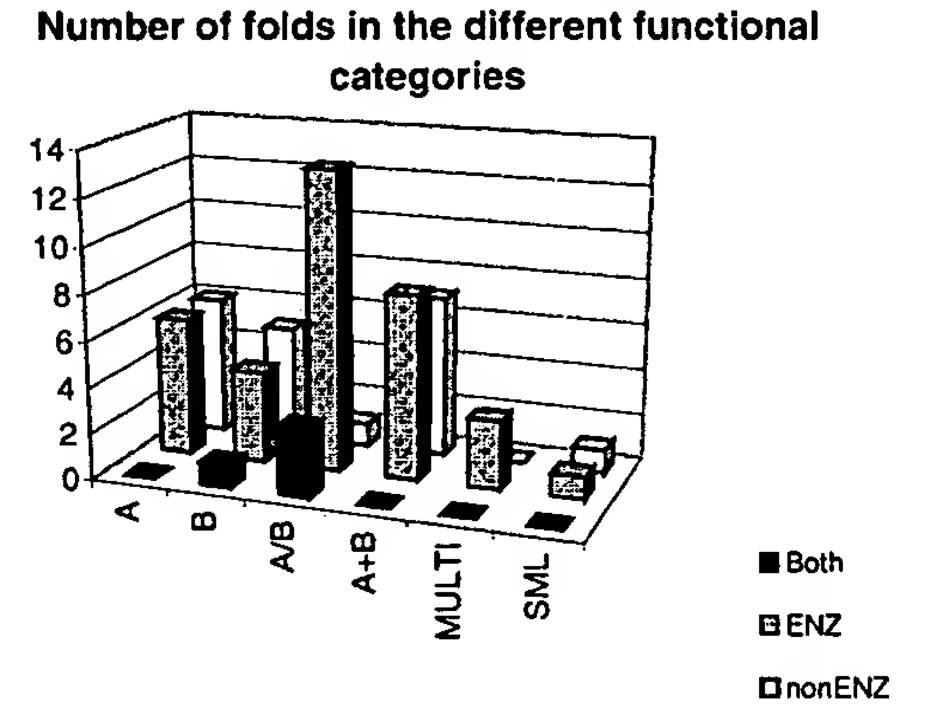
The COGs system classifies functions for those proteins that have clear orthologues in different species. Thus, conclusions based on using yeast COGs should be readily applicable to other genomes. This point is highlighted in Figure 4(d), which shows a COGs *versus* SCOP classification for only the 110 COGs that are conserved across all the analyzed genomes (eight) and all three kingdoms. Thus, this sub-figure would appear *exactly* the same for *E. coli*, *Methanococcos jannaschii* or a number of other genomes. It clearly shows how much more common the information processing proteins are among the most conserved and ancient proteins. Moreover, note how these most ancient proteins appear to have less of a preference for a particular structural class than the "more modern" metabolic ones. This suggests that large-scale duplication of alpha/beta folds for use in metabolism is what gave rise to stronger fold-function association in Figure 3(c).

A. All of Swissprot



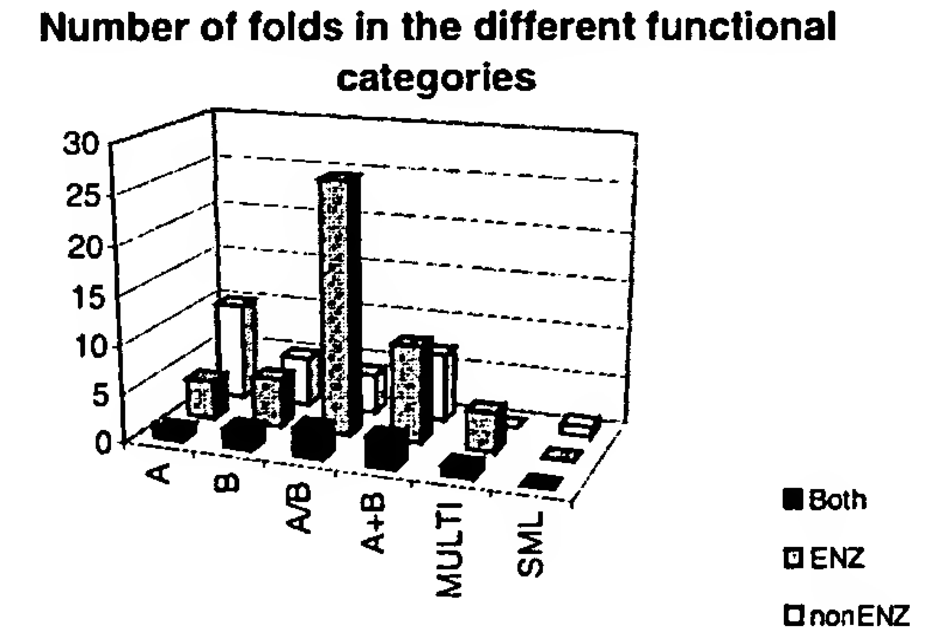
	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	4	13	9	6	2	1	35
ENZ	12	6	34	28	11	2	93
nonENZ	30	17	5	22	2	25	101

B. Yeast



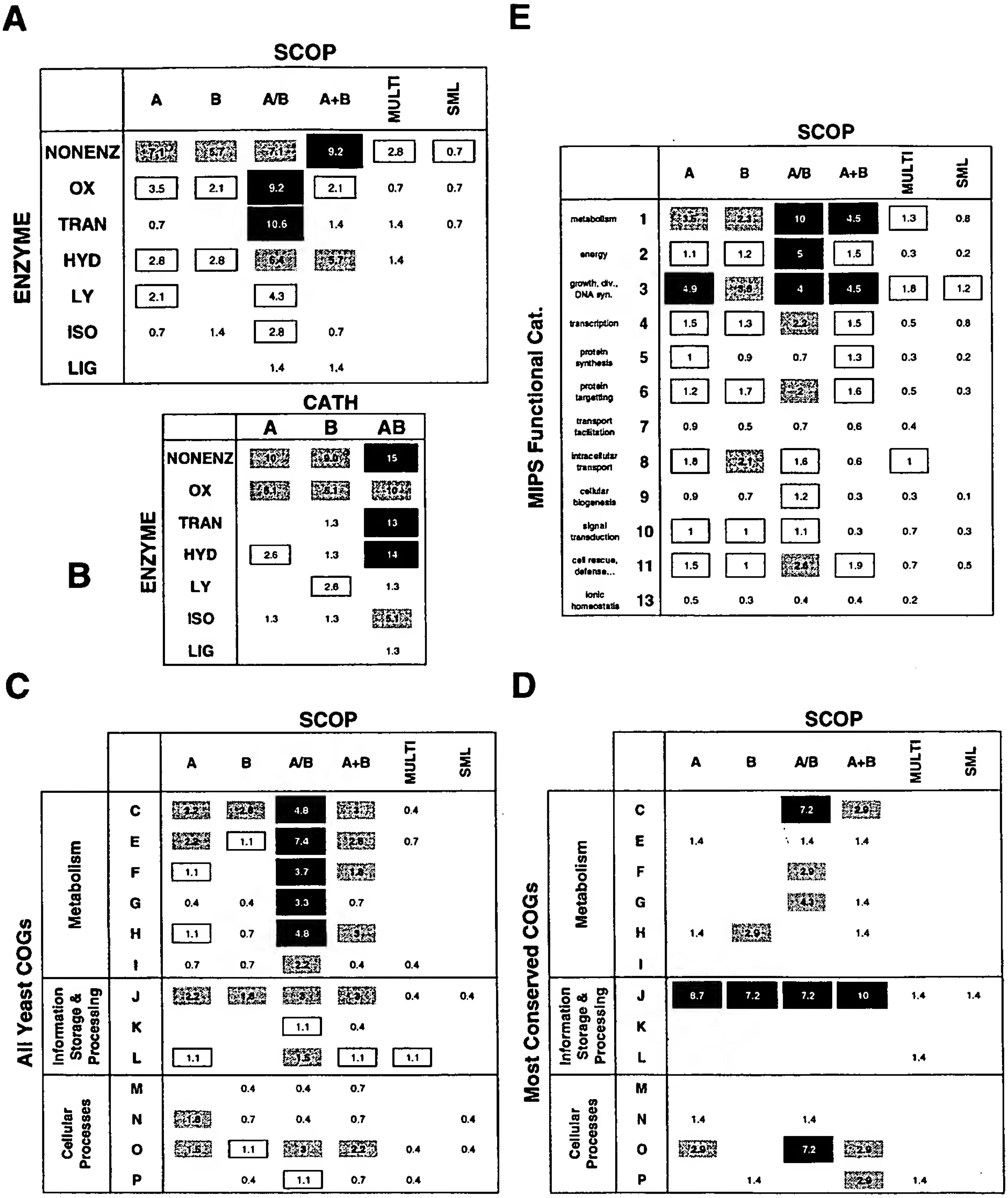
	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	0	1	3	0	0	0	4
ENZ	6	4	13	8	3	1	35
nonENZ	6	5	1	7	0	1	20

C. E. coli



	A	B	A/B	A+B	MULTI	SML	TOTAL
Both	1	2	3	3	1	0	10
ENZ	4	5	26	10	4	0	49
nonENZ	10	5	4	7	0	1	27

Figure 3. Chart with breakdown among structure-function classes in two genomes. Charts and Tables showing the number of folds in each fold class associated with only enzymatic (ENZ), only non-enzymatic (nonENZ), and both enzymatic and non-enzymatic functions (Both). The results are shown for (a) all of Swissprot, (b) for just the yeast genome, and (c) for just the *E. coli* genome. The results for individual domains in a minimum set of SCOP domains also support these tendencies (data not shown). The numbers in (b) are not based on the PSI-blast protocol used for Figure 4. Rather they are found just as "subsets" of the overall Swissprot results to make them readily comparable with the rest of the paper. Because of this the numbers in this Figure will not match exactly those in Figure 4, the difference having to do with the greater number of fold-function combinations found by PSI-blast as compared to WU-blast.



A

SCOP

ENZYME	A	B	A/B	A+B	MULTI	SML
NONENZ	7.1	5.7	7.1	9.2	2.8	0.7
OX	3.5	2.1	9.2	2.1	0.7	0.7
TRAN	0.7		10.6	1.4	1.4	0.7
HYD	2.8	2.8	6.4	5.7	1.4	
LY	2.1		4.3			
ISO	0.7	1.4	2.8	0.7		
LIG			1.4	1.4		

B

CATH

ENZYME	A	B	AB
NONENZ	10	9.0	15
OX	6.1	5.1	10
TRAN		1.3	13
HYD	2.6	1.3	14
LY		2.6	1.3
ISO	1.3	1.3	5.1
LIG			1.3

C

SCOP

All Yeast COGs		A	B	A/B	A+B	MULTI	SML
Metabolism	C	2.2	2.6	4.8	3	0.4	
	E	2.2	1.1	7.4	2.6	0.7	
	F	1.1		3.7	1.8		
	G	0.4	0.4	3.3	0.7		
	H	1.1	0.7	4.8	3		
	I	0.7	0.7	2.2	0.4	0.4	
Information Storage & Processing	J	2.2	1.8	3	3	0.4	0.4
	K			1.1	0.4		
	L	1.1		1.5	1.1	1.1	
Cellular Processes	M		0.4	0.4	0.7		
	N	1.8	0.7	0.4	0.7		0.4
	O	1.5	1.1	3	2.2	0.4	0.4
	P		0.4	1.1	0.7	0.4	

D

SCOP

Most Conserved COGs		A	B	A/B	A+B	MULTI	SML
Metabolism	C			7.2	2.9		
	E	1.4		1.4	1.4		
	F			2.9			
	G			4.3	1.4		
	H	1.4	2.9		1.4		
	I						
Information Storage & Processing	J	8.7	7.2	7.2	10	1.4	1.4
	K						
	L					1.4	
Cellular Processes	M						
	N	1.4		1.4			
	O	2.9		7.2	2.9		
	P		1.4		2.9	1.4	

E

SCOP

MIPS Functional Cat.		A	B	A/B	A+B	MULTI	SML
metabolism	1	3.5	2.3	10	4.5	1.3	0.8
energy	2	1.1	1.2	5	1.5	0.3	0.2
growth, div., DNA syn.	3	4.9	3.8	4	4.5	1.8	1.2
transcription	4	1.5	1.3	2.2	1.5	0.5	0.8
protein synthesis	5	1	0.9	0.7	1.3	0.3	0.2
protein targetting	6	1.2	1.7	2	1.6	0.5	0.3
transport facilitation	7	0.9	0.5	0.7	0.6	0.4	
intracellular transport	8	1.8	2.1	1.6	0.6	1	
cellular biogenesis	9	0.9	0.7	1.2	0.3	0.3	0.1
signal transduction	10	1	1	1.1	0.3	0.7	0.3
cell rescue, defense...	11	1.5	1	2.6	1.9	0.7	0.5
ionic homeostatis	13	0.5	0.3	0.4	0.4	0.2	

Figure 4. Structure-function classes in the yeast genome analyzed through a variety of classification schemes. This Figure shows the distribution of fold function combinations in the yeast genome as analyzed by a variety of different structure and functional classifications. Each of the Figures is a cross-tabulation of one structural classification scheme (on the column heads) *versus* a functional classification (row heads). (a) SCOP *versus* ENZYME; (b) CATH *versus* ENZYME; (c) SCOP *versus* COGs; (d) SCOP *versus* Most Conversed COGs; (e) SCOP *versus* MIPS Functional Catalogue. Each of the grid boxes gives the number of fold-function combinations within a structure-function class. This number is expressed as a percentage of the total number of combinations in the diagram to make the graphs readily comparable. The total number of combinations in each of the sub-figures is (a) 141, (b) 77, (c) 1207, (d) 120, and (e) 66. (a) and (e) are directly comparable with the cross tabulation in Table 2B for all of Swissprot. In (d) and (e), we employ the COGs scheme in exactly the same fashion as we did the ENZYME classification. We form combinations between individual yeast COGs and SCOP folds (e.g. COG 0186 with fold 2.26) and then we place these combinations into larger structure-function classes. The COGs overall functional classes are denoted by a single letter and then are in turn grouped into three broader areas (so, for instance, the 0186-2.26 pair would go into the structure-function class all-beta, J). We, likewise, proceed similarly for the MIPS yeast functional catalogue. This assigns to each function a two or three component number similar to an EC number (e.g. 07.20.3 or 06.2). We use the first two numbers to create combinations with SCOP folds and then use the top number to create the functional classes shown in the diagram. For (e) we just use the 110 COGs that are present in all eight genomes in the current COGs analysis (*E. coli*, *H. influenzae*, *H. pylori*, *M. genitalium*, *M. pneumoniae*, *Synechocystis*, *M. jannaschii*, and yeast).

Top Multifunctional Folds→

		16	9	6	6	6	5	4	4	4	3	3	3	3	3	3	3
		1byb	2ace	1xel	1gky	1fxd	1pho	3chy	1ama	1bdo	1jbc	1snc	1lxa	3pte	1lmd	1fha	1rie
		3.001	3.048	3.018	3.024	4.031	1.063	3.013	3.045	2.055	2.018	2.024	2.053	5.003	5.007	1.021	7.035
NONENZ	0.0.0	7	3	14	295	119		60		11	75	68	1	1	7	42	
OX	1.1.1	22		107													
	1.1.3	2					1										
	1.10.2																5
	1.10.99																1
	1.11.1		4														
	1.14.13						3									6	
	1.14.14	13					44										
	1.14.15						2										
	1.14.99						5										
	1.17.4															38	
	1.18.6				42												
	1.3.1	3		22	3												
	1.3.99	2															
	1.6.5					2											
	1.6.99	7		2				4									
	1.9.3																6
TRAN	2.1.3					3				1							
	2.3.1		6										8				
	2.6.1								65								
	2.7.1									5							
	2.7.4				57	39											
	2.7.7														1		
HYD	3.1.1		56					12			1						
	3.1.2		1														
	3.1.3														36		
	3.1.31											4					
	3.1.4	4															
	3.2.1	98									33						
	3.2.3	3															
	3.4.11		2														
	3.4.16		4											1			
	3.5.2													52			
	3.5.4	5															
	3.6.1					14						19					
	3.7.1		2														
	3.8.1		3														
LY	4.1.1	14							1								
	4.1.2	29															
	4.1.3	2						1									
	4.1.99								7								
	4.2.1	38		15									1				
ISO	5.1.3			25													
	5.3.1	80															
	5.3.3					1											
	5.4.3								1								
	5.4.99							1									
LIG	6.3.3				9												
	6.3.4				17												
	6.4.1									5							

Figure 5. The most versatile folds. The functions associated with the 16 most versatile folds are shown. Values in the table denote the number of matches between a particular fold type in pdb95d (designated by its fold number in SCOP 1.35) and an enzyme category (represented by the first three components of the respective EC numbers). Here and in the following Tables the same parameters were used for matching as in Figure 2. The numbers in the top row indicate the number of functions a particular fold is associated with. The identifiers above the fold numbers are either PDB or SCOP identifiers of representative structures (the latter only if the PDB entry contains more than one domain or chain). (See the legend to Table 3 for the syntax of SCOP identifiers.) The first row in the table with the artificial 0.0.0 EC number shows the number of matches with non-enzymatic functions. Among the two all-alpha folds in the table, cytochrome P450 (1.063) is exclusively enzymatic, associated with five different enzyme functions, all related to cytochrome P450. Only one alpha + beta fold, ferredoxin (4.031), is present in the table, predominantly with matches with non-enzymatic ferredoxins, but also with enzymes in four different enzyme classes. In the multi-domain class, beta-lactamase/D-ala carboxypeptidase (5.003) has the most matches with penicillinase (EC number 3.5.2) and only one match with a non-enzyme, which also binds penicillin but has no enzymatic activity (Coque *et al.*, 1993). The class of small domains is represented only with one fold, membrane-bound rubredoxin-like (7.035), and has matches only with enzymes. It is possible that some proteins classified as "non-enzymes" may indeed be enzymes, missing the corresponding EC number. In this case, our analysis may be potentially useful in pointing to which non-enzymes may actually be enzymes.

Figure 4(e) shows another functional classification scheme, the MIPS Yeast functional catalogue (Mewes *et al.*, 1997). Unlike the COGs scheme, this has the advantage of being applicable to every yeast open reading frame (ORF). However, it has many more categories and about a third of the yeast ORFs are classified into multiple categories (sometimes five or more), making interpretation of the results a bit more ambiguous.

The most versatile folds and the most versatile functions

Returning to considerations of all of Swissprot, Figure 5 lists the 16 most versatile folds. The top five are the TIM-barrel, the alpha-beta hydrolase fold, the Rossmann fold, the P-loop containing NTP hydrolase fold, and the ferredoxin fold. Four of these are alpha/beta folds and one is alpha + beta. All five have non-enzymatic functions as well as five to 15 enzymatic ones. The most versatile folds include four all-beta and two all-alpha folds.

Figure 6 lists the 18 functions that have the most different folds associated with them, each having at least three associated folds. The most versatile functions are those of glycosidases and carboxylyases (3.2.1 and 4.2.1), which are associated with seven different fold types each, recruited from at least three different fold classes. The next two most versatile functions, the phosphoric monoester hydrolases and the linear monoester hydrolases (3.1.3 and 3.5.1), are associated with six different fold types each. Most of the versatile functions are associated with folds in completely different fold classes. This suggests that these enzymes developed independently, providing many examples of convergent evolution. In contrast, only three functions, all oxidoreductases, are associated with folds in a single class (last three rows in Figure 6). These folds are all alpha/beta, namely the TIM-barrel, Rossmann, and flavodoxin folds.

Specific functional convergences involving different folds

Even on the level of specificity of four-component EC numbers, several enzymatic functions are performed by unrelated structures. Figure 1 shows a dramatic example, two different carbonic anhydrases with the same EC number 4.2.1.1, but with clearly different structures (Kisker *et al.*, 1996). Table 3 shows further examples in a more systematic fashion. Most of these occur in different evolutionary lineages. For instance, the all-alpha vanadium chloroperoxidase occurs only in fungi, while the alpha/beta non-heme chloroperoxidase occurs only in prokaryotes. Another example is beta-glucanase. It has as many as three different structural representations, from three different fold classes. While it has an all-beta structure in *Bacillus subtilis*, it has an all-

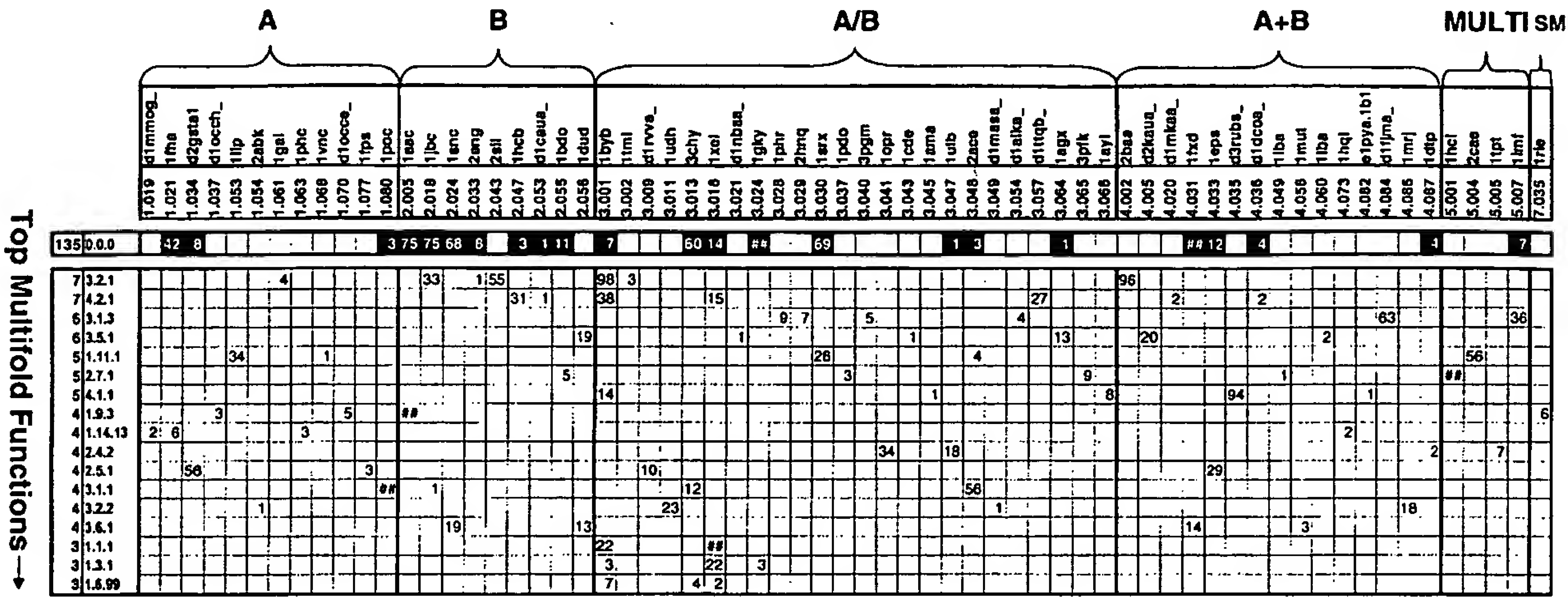


Figure 6. The most versatile functions. Values in the table denote the number of matches between a particular enzyme category (designated by the first three components of their EC numbers) and a SCOP 1.35 fold (designated by their fold numbers). This Figure follows the same conventions described in the legend to Figure 5. The rows are arranged in decreasing order according to the number of different folds with which they are associated (numbers shown in the first column). A hash (#) in any cell indicates that its value is greater than 99.

alpha variant in *Bacillus circulans*, and an alpha/beta structure in tobacco.

Specific functional divergences on same fold

Quite a number of SCOP domains each have sequence similarity with Swissprot proteins of different function. We separated these into cases in which the structural domain has similarity to proteins with different enzymatic functions only and those in which a domain shows homology to both enzymes and non-enzymes (Table 4A and B, respectively). Table 4A includes the well-known lactalbumin-lysozyme C similarity and the well-documented case of homology between an eye-lens structural protein and an enzyme (crystallin and gluthathione *S*-transferase; Cooper *et al.*, 1993; Qasba & Kumar, 1997). It includes several other notable divergences, such as the one between lysophospholipidase and galectin, and the one between an elastase and an antimicrobial protein (Morgan *et al.*, 1991). Remarkably, of the seven domains in this Table, three belong to the all-beta class.

"Multifunctionality" *versus e*-value

Figure 7 shows how the number of "multifunctional" domains, i.e. domains with sequence similarity to proteins with different functions, varies as the function of the stringency of the match score threshold. We used a minimal version of SCOP in which the structures in PDB were clustered into 990 representative domains (see the legend to Figure 7). The Figure shows how the percentage of domains that have sequence similarity to proteins

Table 3. Specific convergences

EC #	Enzymatic function	Fold #1	Dom #1	Swissprot 1	Fold #2	Dom #2	Swissprot 2
1.11.1.10	Chloroperoxidase	3.048.001	d1broa_	PRXC_PSEPY	1.068.001	d1vnc__	PRXC_CURIN
1.15.1.1	Superoxide dismutase	2.001.007	d1srda_	SOD1_ORYSA	4.023.001	d1mnga2	SODM_BACCA
3.1.3.48	Protein-tyrosine phosphatase	3.028.001	d1phr__	PTPA_STRCO	3.029.001	d2hnp__	PYP3_SCHPO
3.1.26.4	Ribonuclease h	3.038.003	d2rn2__	RNH_ECOLI	3.039.001	d1tfr__	RNH_BPT4
3.2.1.4	Endoglucanase	1.061.001	d1cem__	GUN_BACSP	3.001.001	d1ecea_	GUN_BACPO
3.2.1.8	Xylanase	2.018.001	d1yna__	XYN_TRIHA	3.001.001	d2exo__	XYNB_THENE
3.2.1.14	Endochitinase	3.001.001	d1hvq__	CHIA_TOBAC	4.002.001	d2baa__	CHIX_PEA
3.2.1.73	Beta-glucanase*	3.001.001	d1ghr__	GUB_NICPL	2.018.001	d1gbg__	GUB_BACSU
3.2.1.73	Beta-glucanase	1.061.001	d1cem__	GUB_BACCI			
3.2.1.91	Exoglucanase	2.018.001	d1cela_	GUX1_TRIVI	3.002.001	d1cb2a_	GUX3_AGABI
3.5.2.6	Beta-lactamase	5.003.001	d1btl__	BLP4_PSEAE	4.083.001	d1bmc__	BLAB_BACCE
4.2.1.1	Carbonic anhydrase	2.053.001	d1thja_	CAH_METTE	2.047.001	d2cba__	CAHZ_BRARE
5.2.1.8	*Cis-trans* isomerase	4.018.001	d1fkd__	MIP_TRYCR	2.041.001	d2cpl__	CYPR_DROME
5.4.99.5	Chorismate mutase	1.079.001	d1csma_	CHMU_YEAST	4.037.001	d2chsa_	CHMU_BACSU

Explicit enzymatic functions associated with different folds. Of the 13 different enzyme functions listed, eight are hydrolases, five of which belong to the 3.2.1 EC category. One of them, beta-glucanase, is associated with three different folds. Note that most of the enzymes in the Table are associated with folds from different classes. Even when the folds are from the same class, as in the case of protein-tyrosine phosphatases, they are clearly different. Fold numbers are from SCOP 1.35. Domain identifiers are according to the scop syntax: d1pdbc*N*, where "1pdb" is a PDB code, *c* is a chain identifier, and *N* describes if this is the first, second, or only domain in the chain. Thus, d1ggta1 is the first domain in the A chain of 1GGT.

Table 4. Specific divergences

A. *Two different enzymatic functions*

SCOP domain	Fold number	Swissprot 1	EC num 1	Function 1	Swissprot 2	EC num 2	Function 2
d2abk__	1.001.054.001.001.001	END3_ECOLI	4.2.99.18	Endonuclease III	GTMR_METTF	3.2.2.-	Possible G-T mismatches repair enzyme
d1bdo__	1.002.055.001.001.001	BCCP_ECOLI	6.4.1.2	Biotin carboxyl carrier protein of acetyl-Coa carboxylase	BCCP_PROFR	2.1.3.1	Biotin carboxyl carrier protein of methylmalonyl-CoA carboxyl-transferase
d1dhpa_	1.003.001.003.001.004	NPL_ECOLI	4.1.3.3	*N*-Acetylneuraminate lyase subunit	DAPA_BACSU	4.2.1.52	Dihydrodipicolinate synthase
d1hdca_	1.003.018.001.002.005	ENTA_ECOLI	1.3.1.28	2,3 Dihydro-2,3 dihydroxy-benzoate dehydrogenase	ADHI_DROMO	1.1.1.1	Alcohol dehydrogenase 1
d1nipa_	1.003.024.001.005.003	BCHL_RHOCA	1.3.1.33	Protochlorophillide reductase 33 kD subunit	NIFH_THIFE	1.18.6.1	Nitrogenase iron protein
d1gara_	1.003.043.001.001.001	PUR3_YEAST	2.1.2.2	Phosphoribosylglycinamide formyltransferase	PURU_CORSP	3.5.1.10	Formyltetrahydrofolate deformylase
d2dkb__	1.003.045.001.003.001	OAT_RAT	2.6.1.13	Ornithine aminotransferase precursor	GSAB_BACSU	5.4.3.8	Glutamate-1-semialdehyde 2,1-aminomutase 2
d1ede__	1.003.048.001.003.001	DMPD_PSEPU	3.1.1.-	2-Hydroxymuconic semialdehyde hydrolase	HALO_XANAU	3.8.1.5	Haloalkane dehalogenase
d1fua__	1.003.053.001.001.001	ARAD_ECOLI	5.1.3.4	L-Ribulose-5-phosphate 4-epimerase	FUCA_ECOLI	4.1.2.17	L-Fuculose phosphate aldolase
d1lmn__	1.004.002.001.002.010	LCA_RAT	2.4.1.22	Alpha-lactalbumin precursor	LYC1_PIG	3.2.1.17	Lysozyme C-1
d1frva_	1.005.015.001.001.001	FRHG_METVO	1.12.99.1	Coenzyme F420 hydrogenase gamma subunit	MBHS_AZOCH	1.18.99.1	Uptake hydrogenase small subunit precursor

B. *Enzyme and non-enzyme*

SCOP domain	Fold number	Swissprot 1	Enzymatic function	EC number	Swissprot 2	Non-enzymatic function
d1gsq_1	1.001.034.001.001.007	GTS2_MANSE	Glutathione *S*-transferase 2	2.5.1.18	SC11_OMMSL	*S*-Crystallin SL11 (major lens polypeptide)
d1lcl__	1.002.018.001.003.003	LPPL_HUMAN	Eosinophil lysophospholipase	3.1.1.5	LEG7_RAT	Galectin-7
d1brbe_	1.002.029.001.002.003	CFAD_RAT	Endogenous vascular elastase	3.4.21.46	CAP7_HUMAN	Azurocidin (antimicrobial, heparin-binding protein)
d1mup__ ..	1.002.039.001.001.007	PGHD_HUMAN	Prostaglandin-D synthase	5.3.99.2	LACC_CANFA	Beta-lactoglobulin III
..d1mup__	1.002.039.001.001.007				QSP_CHICK	Quiescence-specific protein
d2hhma_ ..	1.005.007.001.002.001	MYOP_XENLA	Inositol mono-phosphatase	3.1.3.25	SUHB_ECOLI	Extragenic suppressor protein SUHB
..d2hhma_	1.005.007.001.002.001	STRO_STRGR	DTDP-glucose synthase	2.7.7.24		
d1isua_	1.007.029.001.001.001	IRO_THIFE	Iron oxidase precursor (FE(II) oxidase)	1.16.3.-	HPIT_RHOTE	High potential iron-sulfur protein (HIPIP)

List of SCOP domains that are each homologous to several Swissprot proteins with significantly different function. In A, the domains homologous to proteins with different (in the last three component of EC numbers) enzymatic functions are listed. In most cases, the enzymatic functions remain analogous, as reflected in the names of the enzymes. B lists the domains homologous to proteins with both enzymatic and non-enzymatic functions. (See Table 3 for the SCOP domain syntax.)

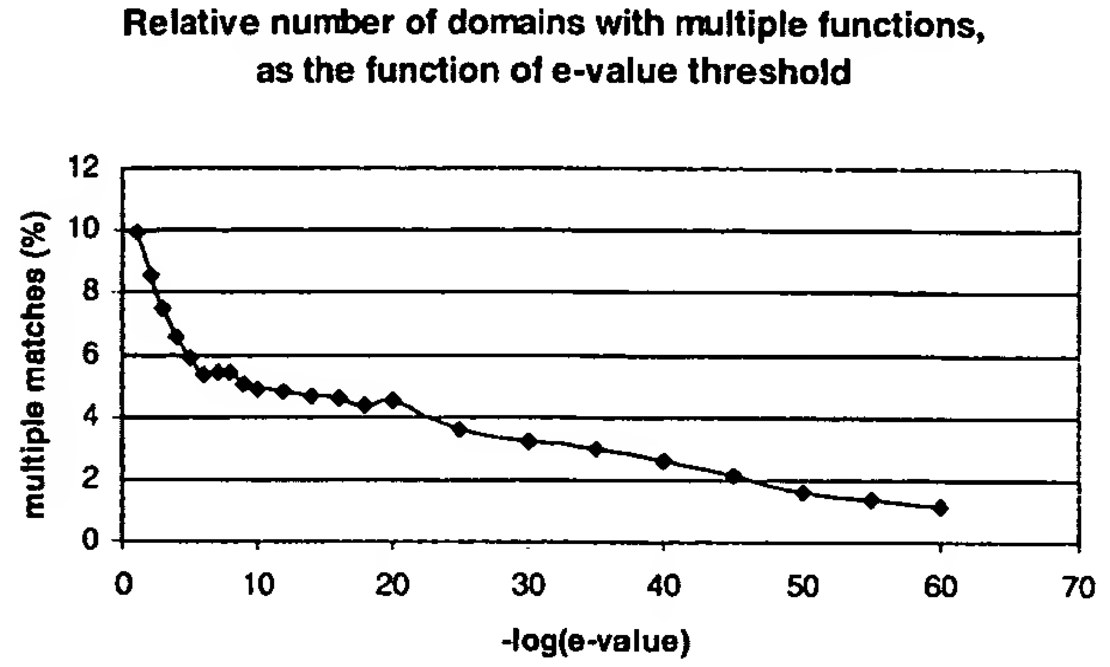


Figure 7. Multi-functionality *versus* *e*-value threshold. The graph shows how the percentage number of multi-functional enzymatic domains varies as the function of the *e*-value threshold. A multi-functional domain occurs when a particular domain in SCOP matches domains in Swissprot with different enzymatic function. For these calculations, we had to use a more minimal version of SCOP than the pdb95d dataset referred to in the methods to prevent double matches, i.e. two SCOP domains matching a single Swissprot domain. The construction of this minimal SCOP was described previously (Gerstein, 1998a). Basically, all the domains in SCOP were clustered *via* a multi-linkage approach into 990 representative domains, such that no two domains matched each other with a FastA *e*-value better than 0.01.

with different functions (in terms of three-component EC numbers) varies with sequence similarity. This decreases approximately monotonically as a function of the exponent of the *e*-value threshold. Interestingly, there is a breaking point around log (*e*-value) = −5, as the sharply decreasing number of functions slows down and the matches reach the level of biological significance.

Our graph can be loosely compared with the classic graph by Chothia & Lesk (1986) showing the relation of similarity in structure to that in sequence. It roughly shows the chance of functional similarity (or more precisely the chance of functional difference) with a given level of sequence similarity between an enzyme and a protein of unknown function. For example, with an *e*-value of 10^{-10}, there is only an ~5% chance that an unknown protein homologous to a certain enzyme has in fact a different function. Moreover, our graph is in excellent agreement with the findings by Russell *et al.* (1998) who also found that the proportion of homologues with different functions is around 10%. This shows that there is a low chance that a single-domain protein, highly homologous to a known enzyme, has a different function.

Discussion and Conclusions

Overview

We have investigated the relationship between the structure and function of proteins by comparing functionally characterized enzymes in Swissprot with structurally characterized domains in SCOP. It is a timely subject, as the number of three-dimensional protein structures is increasing rapidly and the recent completion of several microbial genomes highlights the need for functional characterization of the gene products and identification of enzymes participating in metabolic pathways (Koonin *et al.*, 1998).

We tried to be as objective and as unbiased as possible, taking only enzymes with a single assigned function and only single-domain matches. We ignored Swissprot proteins with dubious or unknown function, or with incomplete sequence. Given these criteria, several tendencies are clear. The alpha/beta folds tend to be enzymes. The all-alpha folds tend to be non-enzymes and the all-beta and alpha + beta folds tend to have a more even distribution between enzymes and non-enzymes.

Our analysis of proteins from yeast and *E. coli* has shown that the functional distribution of folds does not differ greatly from the whole of Swissprot. *E. coli*, however, appears to have somewhat more alpha/beta enzymes and less non-enzymes.

Functional assignment complexities

We identified four specific complexities in our functional assignment worth mentioning.

Firstly, there is not always a one-to-one relationship between gene protein and reaction (Riley, 1998). An enzyme can have two functions, or two polypeptides from two different genes can oligomerize to perform a single function. It might be that some of the fold-functions combinations in Figure 2 occur together in multi-domain proteins (which otherwise were not the subject of this survey). An exhaustive screening revealed that only four pairs of folds in Figure 2 were present concurrently in multi-domain proteins. Each of these reduced by one the number of independent fold-function combinations. (The four pairs were as follows, with one representative Swissprot protein in each category, EC numbers in parentheses, and then SCOP fold numbers: PTAA_ECOLI (2.7.1) has 4.049 and 2.055 folds, TRP_COPCI (4.2.1) has 3.057 and 4.005 folds, URE1_HELFE (3.5.1) has 4.005 and 2.056 folds, while XYNA_RUMFL (3.2.1) has 2.018 and 3.001 folds.)

Secondly, the functions associated with similar structures often turn out to be analogous, even if they show significant difference in their EC numbers. For example, acetyl-CoA carboxylase and methylmalonyl-CoA carboxyltransferase enzymes are both actually part of enzyme complexes in which they perform the same function, acting as enzyme carriers. This similarity is not reflected in their EC classification numbers (6.4.1.2 and 2.1.3.1, respectively).

Thirdly, there are clearly some drawbacks to the EC system. The EC system is a classification of reactions, not underlying biochemical mechanisms. An enzyme classification system based explicitly on reaction mechanism (e.g. "involves pyridoxal phosphate" or "involves Ser as a nucleophile") might also prove interesting to compare with protein structure. Alternatively, one based on pathways might be worthwhile since, as pointed out by Martin *et al.* (1998), "it may be that more significant relationships occur within pathways, where the substrate is successively transferred from enzyme to enzyme along the pathway, requiring similar binding sites at each stage".

Finally, in all of Swissprot the majority of the 101 folds with only non-enzymatic functions probably have several functions, but we were not able to consider them separately here, lacking a general protein function classification system for non-enzymes. Such a system is not easy to derive. For instance, if we took only the first three words of all the description lines in Swissprot, we would end up with about 10,000 different protein functions (besides enzymes). An approximate solution to this problem is offered by a recent work that has classified 81% of Swissprot into one of three broad categories in an automated fashion (Tamames *et al.*, 1997). However, one way we did tackle this problem was by focussing on the yeast genome for which there are a number of overall functional classification systems. This work showed that the preferred association of folds with certain functions occurs for non-enzymes as well as enzymes. Furthermore, the results for the highly conserved COGs would be expected to be exactly the same in other genomes.

Biases

Our results are undoubtedly affected to some degree by the biases inherent in the databanks, e.g. towards mammalian, medically relevant proteins and towards proteins that easily crystallize. Such biases probably result in the higher representation of enzymes in the structural databases, in the PDB and therefore in SCOP. This might be the cause of the higher occurrence of alpha/beta proteins in our tables and the higher density of matches in this class.

One interesting question related to biases is whether looking only at individual genomes instead of the whole database will give different results. Our results for yeast suggest that it is not necessarily the case.

Comparison with Martin *et al.* (1998)

Martin *et al.* (1998) performed a similar analysis to the one described here. One of the conclusions of their careful study was that there was no relationship between the top-level CATH classification and the top-level EC class. This seems to be at odds with our results. However, we have found the conclusions to be consistent. There are a number of reasons for this.

Firstly, Martin *et al.* (1998) tabulate statistics on only the proteins in the PDB. They found a clear alpha/beta preference for proteins in the oxidoreductase, transferase, and hydrolase categories (EC 1-3), but for the lyase, isomerase, and ligase categories (EC 4-6) they observe different tendencies. However, they did not have sufficient counts to establish statistical significance for this latter finding. (This is basically what we observe in Figure 4(b).) Because in our analysis we use all of Swissprot and we tabulate our statistics a little differently (in terms of combinations), we get more "counts" than Martin *et al.* (1998). Thus, we are able to argue that the different distribution of fold-function combinations observed for lyases, isomerases, and ligases are significant. This is borne out by the chi-squared statistics at the end of Table 2.

Secondly, Martin *et al.* "no-relationship" conclusion applies only to comparisons between the different enzyme classes. However, we find our largest differences when comparing non-enzymes to enzymes and also comparing between the various types of non-enzymes.

Finally, the CATH classification that Martin *et al.* use has only three classes in its top-most level. In contrast, SCOP has six top classes (Table 1). While this larger number of categories does tend to degrade our statistics somewhat, it also highlights some differences that cannot be observed in terms of the CATH classes alone, e.g. we find clear differences between alpha + beta and alpha/beta proteins and also between small proteins and all others.

Apparently high occurrence of convergent evolution

Note that the table in Figure 2 is not square: it has more folds than functions. This shape leads to a number of interesting conclusions. The 331 fold-function combinations we observe for 229 folds and 92 functions imply that there are 1.2 functions per fold and 3.6 folds per function. However, these numbers are somewhat skewed by the large number of folds (101) associated only with the single non-enzymatic function. If we exclude these, we get 128 "enzyme-related" folds, which are, in turn, associated with 230 (= 331 − 101) different fold-function combinations. This implies that for the enzyme-related folds there are on average 1.8 functions per fold and 2.5 folds per function (230/128 and 230/92). The larger number of folds per function than functions per fold seems to suggest that nature tends to reinvent an enzymatic function (i.e. convergent evolution) more often than modify an already existing one (i.e. functional divergence).

How can we explain this? Firstly, 1.8 is a lower estimation for the number of functions per fold as the non-enzymatic functions were bundled into one group here. Secondly, there are several examples of functional divergence for a fold within one three-component enzyme category that are not reflected in our Tables. For instance, the 1.1.1 category has 248 different enzymes, which all share the same fold. Thirdly, the results in this paper were derived from databases comprised of data from several organisms. It is quite possible that within one organism, functional divergence is more prevalent than convergent evolution.

Superfolds and superfunctions

Are functions more diverse for the more common folds? To some degree this brings up a "chicken-and-egg" issue. Do folds have more functions because they occur more often or is it the other way around? The commonness of a fold is often quantified by the number of non-homologous sequence families accommodated by the fold, and folds accommodating many families of diverse sequences have been dubbed "superfolds" (Orengo *et al.*, 1993). We find that there seems to be a loose connection between the number of diverse sequence families associated with a particular fold (in SCOP) and the functional diversity of that fold. For instance, the top superfold is the TIM-barrel; it also has the most functions associated with it (15 different enzymatic functions as shown in Figure 4). On the other hand, there are exceptions: the alpha/beta hydrolases and the Rossmann fold are both associated with 22 sequence families in SCOP, but while the former has eight different enzymatic functions, the latter has only three.

Finally, while there is a high incidence of particular functions with many folds ("superfunctions"), as well as folds with many functions, the distribution of superfunctions appears to be more uniform and less concentrated on a few exceptionally versatile individuals than is the case for folds. That is, comparing Figures 3 and 4 one can see that the top nine most versatile functions are associated with five to seven folds while the top nine most versatile folds carry out from six to as many as 16 functions. This last value is for the TIM-barrel and underscores the uniqueness of this fold as a generic scaffold (see Figure 1 for an illustration of this fold).

Why folds are associated with functions: chemistry *versus* history

Why is a certain fold chosen to carry out a particular function? It is, of course not possible to answer this question definitively at present. However, there are two broad themes that emerge from our analysis. The first is favorable chemistry. Perhaps the TIM-barrel design simply provides a "more efficient" scaffold for enzyme reactions so that is why it is so prevalent. Another factor is history. Perhaps the association between a particular fold and its function reflects a particular "accident" that took place at the beginning of cellular evolution. However, once this choice was made it was impossible to undo even if other folds would be more chemically suitable. This could be the situation for the ribosomal proteins (and is borne out by the results of Figure 4(d)).

Materials and Methods

Sequence matching to swissprot

All the protein sequences in Swissprot 35 were compared with all the protein domain sequences in SCOP 1.35 by standard database search programs (WU-BLAST; Altschul *et al.*, 1990). The following five criteria were used in the searches: (1) At least three of the four components of the EC number are assigned in the DE line of the Swissprot entries. (2) Fragments in Swissprot were excluded (this affected about 10% of the entries). (3) For WU-BLAST searches an *e*-value threshold of 0.0001 was used, unless stated otherwise. (4) Only "monoenzymes", i.e. proteins with only one enzymatic function, were considered. This excluded less than 0.5% of the Swissprot enzymes. (5) Only single-domain matches with Swissprot proteins were taken into consideration. This means those proteins that had a match with a SCOP domain covering most of the Swissprot protein. Specifically, we required that less than 100 amino acid residues be left uncovered in the Swissprot entry by a match. We are aware that this is only an approximation, as there are domains with less than 100 amino acid residues; however, it is considerably less than the average length of a SCOP domain (163 residues) and seems to be a reasonable threshold in an automated approach.

All the searches were repeated using FASTA with an *e*-value threshold of 0.01 (Pearson, 1998; Pearson & Lipman, 1988). The results obtained by the two different comparison programs were in agreement with each other. That is, the FASTA searches did not result in any new combinations of folds and enzymatic functions (a new dot in Figure 1), and therefore are not shown.

Sequence matching to the yeast genome

To get as great a coverage of the yeast genome as possible, we did a sequence comparison for *just* Figure 4 using an altered protocol. We first ran the PDB against the yeast genome using FASTA and kept all matches with a better than 0.01 *e*-value (Pearson, 1998; Pearson & Lipman, 1988). Then, to increase our number of matches further we used the PSI-blast program (Altschul *et al.*, 1997). This program is somewhat more complex to run than FASTA, involving embedding the yeast genome in NRDB and running PDB query sequences against it in an iterative fashion, adding the matches found at each round to a growing profile. We used the PSI-blast parameters adapted from Teichmann *et al.* (1998): an *e*-value threshold of 0.0005 to include matches in the profile and iteration of up to 30 times or to convergence. We did not continuously parse the output and accepted matches at the final iteration that had *E*-value scores better than 0.0001. The number of iteration to convergence varies depending on the PDB domains being run. Runs that take many iterations such as those for the immunoglobulin superfamily take quite a long time (up to 30 minutes on DEC 500 MHz workstation) and create large output files. In total, PSI-blast finds many more matches than either FASTA or WU-BLAST. However, it has problems with certain small and compositionally biased proteins. We used FASTA for these and also tried to remove compositional bias through running the SEG program with standard parameters (Wootton & Federhen, 1996).

How the structural classifications were used: SCOP and CATH

SCOP hierarchically clusters all the domains in the PDB database, assigning a five-component number to each domain (Murzin *et al.*, 1995). The first component in the SCOP numbers denotes the structural class to which the domain in question belongs. The second component of the SCOP numbers designates the fold type of the domain. There are altogether 361 different fold types in SCOP 1.35. The six SCOP classes used in this survey are listed in Table 1B.

In this study, a 95% non-redundant subset of SCOP was used, i.e. all pairs of domains had less than 95% sequence homology. This set is denoted pdb95d and is available from the SCOP website (scop.mrc-lmb.cam.ac.uk). We used version 1.35, which had 2314 protein domains. (The yeast analysis used a more recent version of SCOP, 1.38, which had 3206 domains.)

The CATH classification classifies structures in analogous fashion to SCOP (Orengo *et al.*, 1997). However, the exact structure of the classification is not the same, with an additional architecture level inserted between the top-level class and the fold-level. In our use of the classification, we created a limited mapping table that associated each SCOP domain in pdb95d with its corresponding classification in CATH 1.4. This was not always possible to do unambiguously. As a result, we left out the ambiguous matches from the statistics.

How the functional classifications were used: ENZYME, COGS, and MIPS

The EC numbers of enzymes are composed of four components (Barrett, 1997). (1) The first component shows to which of the six main divisions the enzyme belongs. (2) The second figure indicates the subclass (referring to the donor in oxidoreductases or the group transferred in transferases, or the affected bond in hydrolases, lyases or ligases). (3) The third figure indicates the sub-subclass (e.g. indicating the type of acceptor in oxidoreductases). (4) The fourth figure gives the serial number of the enzyme in its sub-subclass. The six main divisions are listed in Table 1A.

In the analysis of all of Swissprot, when we counted the number of non-enzymatic matches, all the proteins called 'HYPOTHETICAL' and all the proteins having an '-ase' word ending but lacking an EC number in their description were excluded, because of their functional ambiguity. For relating the sequence matches of the yeast genome to the EC system, we used essentially the same criteria as we did for all of Swissprot (see above): single-domain, monoenzyme matches with at least a three-component EC number.

The COGs and especially the MIPS classifications are a bit more complex than the EC system in that they include non-enzymes as well as enzymes (Tatusov *et al.*, 1997; Koonin *et al.*, 1998; Mewes *et al.*, 1997). They often associate multiple functions or roles to a given yeast ORF. This happens for more than a third of the yeast ORFs with MIPS. In this case, if we could clearly show a PDB match was associated with a single functional domain we made only that pairing. Otherwise we associated all the functions assigned to a given PDB match to its respective fold.

Availability of results over the internet

A number of detailed tables relevant to our study will be made available over the Internet at http://bioinfo.mbb.yale.edu/genome/foldfunc, in particular, a "clickable" version of Figure 1 and large data files giving all the fold assignment and fold-function combinations for Swissprot and yeast.

Acknowledgments

We thank the Donaghue Foundation and the ONR for financial support (grant N000149710725). We thank Ted Johnson for help with the minimal version of the SCOP database.

References

Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. & Selley, J. N. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucl. Acids Res.* **26**, 304-308.

Bairoch, A. (1996). The ENZYME data bank in 1995. *Nucl. Acids Res.* **24**, 221-222.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* **26**, 38-42.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217-221.

Barrett, A. J. (1997). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur. J. Biochem.* **250**, 1-6.

Bork, P. & Eisenberg, D. (1998). Deriving biological knowledge from genomic sequences. *Curr. Opin. Struct. Biol.* **8**, 331-332.

Bork, P. & Koonin, E. V. (1998). Predicting functions from protein sequences-where are the bottlenecks? *Nature Genet.* **18**, 313-318.

Bork, P., Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Sci.* **2**, 31-40.

Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393-403.

Chen, L., DeVries, A. L. & Cheng, C. H. (1997). Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl Acad. Sci. USA*, **94**, 3817-3822.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.

Cooper, D. L., Isola, N. R., Stevenson, K. & Baptist, E. W. (1993). Members of the ALDH gene family are lens and corneal crystallins. *Advan. Exp. Med. Biol.* **328**, 169-179.

Coque, J. J., Liras, P. & Martin, J. F. (1993). Genes for a beta-lactamase, a penicillin-binding protein and a transmembrane protein are clustered with the cephamycin biosynthetic genes in *Nocardia lactamdurans*. *EMBO J.* **12**, 631-639.

Corpet, F., Gouzy, J. & Kahn, D. (1998). The ProDom database of protein domain families. *Nucl. Acids Res.* **26**, 323-326.

des, Jardins M., Karp, P. D., Krummenacker, M., Lee, T. J. & Ouzounis, C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB*, **5**, 92-99.

Doolittle, R. F. (1994). Convergent evolution: the need to be explicit. *Trends Biochem. Sci.* **19**, 15-18.

Fabian, P., Murvai, J., Hatsagi, Z., Vlahovicek, K., Hegyi, H. & Pongor, S. (1997). The SBASE protein domain library, release 5.0: a collection of annotated protein sequence segments. *Nucl. Acids Res.* **25**, 240-243.

Frishman, D. & Mewes, H.-W. (1997). Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4**, 626-628.

Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8**, 779-790.

Gerstein, M. (1997). A structural census of genomes: comparing eukaryotic, bacterial and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. (1998a). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Design*, **3**, 497-512.

Gerstein, M. (1998b). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518-534.

Gerstein, M. & Hegyi, H. (1998). Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **22**, 277-304.

Gerstein, M. & Levitt, M. (1997). A structural census of the current population of protein sequences. *Proc. Natl Acad. Sci. USA*, **94**, 11911-11916.

Hellinga, H. W. (1997). Rational protein design: combining theory and experiment. *Proc. Natl Acad. Sci. USA*, **94**, 10015-10017.

Hellinga, H. W. (1998). Computational protein engineering. *Nature Struct. Biol.* **5**, 525-527.

Henikoff, S., Pietrokovski, S. & Henikoff, J. G. (1998). Superior performance in protein homology detection with the Blocks Database servers. *Nucl. Acids Res.* **26**, 309-312.

Hodges, P. E., Payne, W. E. & Garrels, J. I. (1998). The Yeast Protein Database (YPD): a curated proteome database for *Saccharomyces cerevisiae*. *Nucl. Acids Res.* **26**, 68-72.

Holm; L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316-319.

Ibba, M., Bono, J. L., Rosa, P. A. & Soll, D. (1997a). Archaeal-type lysyl-tRNA synthetase in the Lyme disease spirochete *Borrelia burgdorferi*. *Proc. Natl Acad. Sci. USA*, **94**, 14383-14388.

Ibba, M., Morgan, S., Curnow, A. W., Pridmore, D. R., Vothknecht, U. C., Gardner, W., Lin, W., Woese, C. R. & Soll, D. (1997b). A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science*, **278**, 1119-1122.

Karp, P. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics*, **14**, 753-754.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1998). EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **26**, 50-53.

Kisker, C., Schindelin, H., Alber, B. E., Ferry, J. G. & Rees, D. C. (1996). A left-hand beta-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon *Methanosarcina thermophila*. *EMBO J.* **15**, 2323-2330.

Koonin, E. V. & Galperin, M. Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* **7**, 757-763.

Koonin, E. V. & Tatusov, R. L. (1994). Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J. Mol. Biol.* **244**, 125-132.

Koonin, E. V., Tatusov, R. L. & Galperin, M. Y. (1998). Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **8**, 355-363.

Kraulis, P. J. (1991). MOLSCRIPT-a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. & Thornton, J. M. (1998). Protein folds and functions. *Structure*, **6**, 875-884.

Marvin, J. S., Corcoran, E. E., Hattangadi, N. A., Zhang, J. V., Gere, S. A. & Hellinga, H. W. (1997). The rational design of allosteric interactions in a monomeric protein and its applications to the construction of biosensors. *Proc. Natl Acad. Sci. USA*, **94**, 4366-4371.

Mewes, H. W., Albermann, K., Bahr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F. & Zollner, A. (1997). Overview of the yeast genome. *Nature*, **387**, 7-65.

Morgan, J. G., Sukiennicki, T., Pereira, H. A., Spitznagel, J. K., Guerra, M. E. & Larrick, J. W. (1991). Cloning of the cDNA for the serine protease homolog CAP37/azurocidin, a microbicidal and chemotactic protein from human granulocytes. *J. Immunol.* **147**, 3210-3214.

Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **27**, 29-34.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identifying and classifying protein fold families. *Protein Eng.* **6**, 485-500.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH-a hierarchic classification of protein domain structures. *Structure*, **5**, 1093-1108.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266**, 227-259.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71-84.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA*, **85**, 2444-2448.

Qasba, P. K. & Kumar, S. (1997). Molecular divergence of lysozymes and alpha-lactalbumin. *Crit. Rev. Biochem. Mol. Biol.* **32**, 255-306.

Riley, M. (1997). Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucl. Acids Res.* **25**, 51-52.

Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211-1227.

Russell, R. B., Sasieni, P. D. & Sternberg, M. J. E. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.

Seery, L. T., Nestor, P. V. & FitzGerald, G. A. (1998). Molecular evolution of the aldo-keto reductase gene superfamily. *J. Mol. Evol.* **46**, 139-146.

Selkov, E., Galimova, M., Goryanin, I., Gretchkin, Y., Ivanova, N., Komarov, Y., Maltsev, N., Mikhailova, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L. & Selkov, E., Jr (1997). The metabolic pathway collection: an update. *Nucl. Acids Res.* **25**, 37-38.

Sonnhammer, E., Eddy, S. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405-420.

Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66-73.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**, 631-637.

Teichmann, S., Park, J. & Chothia, C. (1998). Structural assignments to the proteins of *Mycoplasma genitalium* show that they have been formed by extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA*, **95**, 14658-14663.

Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554-571.

Edited by G. von Heijne

(Received 16 November 1998; received in revised form 1 March 1999; accepted 1 March 1999)